Universidade de Brasília – UnB
Faculdade UnB Gama – FGA
Engenharia de Software

O aprendizado de máquina como ferramenta de *business intelligence* na melhoria da gestão do paciente

Autor: Izabela Cristina Nere Rodrigues Cardoso
Orientador: Doutora Carla Silva Rocha Aguiar

Brasília, DF
2018

Izabela Cristina Nere Rodrigues Cardoso

O aprendizado de máquina como ferramenta de *business intelligence* na melhoria da gestão do paciente

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Orientador: Doutora Carla Silva Rocha Aguiar

Coorientador: Doutor Paulo Roberto Miranda Meirelles

Brasília, DF

2018

Izabela Cristina Nere Rodrigues Cardoso

O aprendizado de máquina como ferramenta de *business intelligence* na melhoria da gestão do paciente/ Izabela Cristina Nere Rodrigues Cardoso. – Brasília, DF, 2018-

65 p. : il. (algumas color.) ; 30 cm.

Orientador: Doutora Carla Silva Rocha Aguiar

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2018.

1. data science. 2. health. I. Doutora Carla Silva Rocha Aguiar. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. O aprendizado de máquina como ferramenta de *business intelligence* na melhoria da gestão do paciente

CDU 02:141:005.6

Izabela Cristina Nere Rodrigues Cardoso

O aprendizado de máquina como ferramenta de *business intelligence* na melhoria da gestão do paciente

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, 09 de Março de 2018 – Data da aprovação do trabalho:

Doutora Carla Silva Rocha Aguiar
Orientador

Brasília, DF
2018

Resumo

O dado por si só não é conhecimento, entretanto carrega o potencial de conhecimentos úteis serem extraídos dele. Com o avanço da tecnologia, várias áreas que faziam registros em papéis passaram a fazê-los de forma digital, a saúde é uma dessas áreas. Com essa transição, uma grande quantidade de dados digitais passou a ser produzida de forma rápida, aumentando as oportunidades de extração de conhecimentos a partir destes dados. Estes conhecimentos podem ser utilizados no suporte à tomada de decisões na área da saúde, permitindo que sejam feitas decisões mais bem informadas. Este é o propósito do *business intelligence*, que une conceitos, métodos e tecnologias para melhorar tomadas de decisões a partir da análise dos dados. Apesar do *business intelligence* ser amplamente aplicado em contextos industriais, sua aplicação na área da saúde é crescente. Desta forma, este trabalho busca desenvolver um modelo de *business intelligence* para a área da saúde, identificando quais são as estratégias mais adequadas para este contexto através de um estudo de caso, avaliando principalmente a contribuição do aprendizado de máquina como estratégia de *business intelligence*.

Palavras-chaves: *business intelligence. data science.* saúde, aprendizado de máquina.

Abstract

The data by itself is not knowledge, but carries the potential of useful knowledge to be extracted from it. With the advancement of technology, several areas that made records in papers began to do them in digital form, healthcare is one of these areas. With this transition, a large amount of digital data started to be produced quickly, increasing the opportunities for extracting knowledge from these data. This knowledge can be used to support decision making in the healthcare area, allowing for more informed decisions. This is the purpose of business intelligence, which unites concepts, methods and technologies to improve decision making from data analysis. Although business intelligence is widely applied in industrial contexts, its application in health is growing. In this way, this work seeks to develop a business intelligence model for the healthcare area, evaluating the most appropriate strategies for this context, through a case study.

Key-words: business intelligence. data science. healthcare.

Lista de ilustrações

Lista de abreviaturas e siglas

FGA	Faculdade do Gama
UnB	Universidade de Brasília
OLAP	*Online Analytical Processing*
BI	*Business Intelligence*
ONU	Organização das Nações Unidas
VistA	*Veterans Health Information Systems and Technology Architecture*
OpenMRS	*Open Medical Record System*

Sumário

1 Introdução

Este capítulo tem como objetivo justificar a escolha do tema (seção 1.1), apresentar os problemas que tornam válido o estudo deste tema (seção 1.2), apresentar a questão de pesquisa a ser investigada (seção 1.3) e descrever os objetivos a serem alcançados (seção 1.4).

1.1 Justificativa

Em 2016, mais da metade da população vivia em áreas urbanas (World Bank Group, 2016), estimativas da ONU (Organizações das Nações Unidas) preveem que até 2050, este número crescerá para 70%. Segundo Washburn et al. (2009, p. 3), "o rápido crescimento populacional trouxe novos desafios para os serviços e infraestrutura das cidades", que por sua vez, estão se tornando "mais inteligentes", já que para superar estes desafios, há cada vez mais dependência da tecnologia.

Dessa forma, o governo tem investido na tecnologia para equipar suas cidades, uma vez que foram identificadas várias áreas em que as cidades inteligentes desempenharão um papel fundamental (WASHBURN et al., 2009), (SOLANAS et al., 2014). Uma destas áreas é a saúde, segundo Solanas et al. (2014), as variáveis fornecidas pela infraestrutura da cidade inteligente permitem o entendimento do ambiente de vida do cidadão. Estas variáveis produzem um grande volume de dados, conhecido como *big data*. Segundo Raghupathi e Raghupathi (2014), o *big data* na saúde diz respeito a grandes conjuntos de dados de saúde eletrônicos, cujo o tamanho e a complexidade os tornam difíceis de serem gerenciados. Entretanto, estes dados carregam a oportunidade de, através da sua análise, fornecer informações para uma tomada de decisões mais bem informada, permitindo que seja fornecido à população, um melhor e mais adaptado serviço.

Por fim, o *business intelligence* (BI) se mostra uma boa ferramenta para esta análise, uma vez que busca fornecer "informações acionáveis entregues no momento certo, no local certo e na forma correta para auxiliar os tomadores de decisão" (NEGASH, 2004, p. 178), suprindo a necessidade da transformação dos dados de saúde em informação, principalmente em cidades urbanas, de forma que eles possam ser úteis para a gestão da saúde publica, facilitando a tomada de decisões clínicas e administrativas (METTLER; VIMARLUND, 2009).

1.2 Problema de Pesquisa

Segundo Haux (2006), nas últimas décadas houve uma grande mudança na manipulação da informação na área da saúde, uma vez que os dados passaram a ser processados e armazenados computacionalmente ao invés de em papéis, melhorando as oportunidades de uso dos mesmos. Entretanto, é importante que o dado seja usado de forma sistemática para a qualidade e eficiência dos serviços de saúde oferecidos, dessa forma, são necessárias "teorias computacionais e ferramentas que auxiliem humanos a extrair informações úteis (conhecimento) do rápido crescimento do volumes de dados digitais" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 37).

O BI é uma destas ferramentas. Ele consiste em duas atividades primárias: a entrada e a saída dos dados. A entrada dos dados prepara o dado de forma a torná-lo apto ao suporte à decisão e proporciona o acesso adequado a ele, entretanto, após acessá-lo, é necessário analisá-lo de forma a auxiliar nas tomadas de decisões, o que consiste na segunda atividade, a saída de dados, onde o BI é implementado de fato, utilizando os dados disponibilizados na etapa anterior para estas análises.

A mineração de dados é um componente crucial na camada de análise do BI, ela consiste na busca de relacionamentos e padrões distintos em um conjunto de dados (ALNOUKARI; RAZOUK; HANANO, 2016). Segundo Bose e Mahapatra (2001, p.211) "técnicas de aprendizado de máquina são usadas para análise de dados e descobrimento de padrões e, portanto, podem desempenhar um papel fundamental no desenvolvimento de aplicações de mineração de dados", motivando a utilização destas técnicas como suporte ao BI.

Desta forma, este trabalho busca analisar o aprendizado de máquina como estratégia de BI, a fim de identificar qual ou quais modelos melhor sanam o problema no uso de dados digitais na área da saúde, que apesar de possuírem um grande volume, nem sempre são usados de forma sistemática, impedindo o aproveitamento de todo o potencial destes dados no suporte à tomada de decisões. Isto será feito utilizando principalmente dados da Secretaria da Saúde de São Paulo, com foco nos dados relacionados à gestão do paciente.

1.3 Questão de Pesquisa

Este trabalho visa então responder a seguinte questão: Qual ou quais estratégias podem ser utilizadas no *business intelligence* de forma a melhor auxiliar no uso de dados para tomada de decisões na área da saúde?

1.4 Objetivos

Os seguintes objetivos gerais e específicos guiaram este trabalho:

1.4.1 Objetivo Geral

Desenvolver um modelo que utilize o *business intelligence* no contexto da área da saúde, de forma a auxiliar organizações de saúde na tomada de decisões baseada em dados, para melhoria nos serviços prestados à população.

1.4.2 Objetivos Específicos

A fim de atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

- Identificar, dos dados disponíveis, quais são relevantes para a tomada de decisões na área da saúde
- Estabelecer acesso adequado ao dado
- Identificar quais estratégias permitem a extração de informações úteis a partir dos dados, com foco no aprendizado de máquina
- Aplicar as estratégias nos dados selecionados, de modo a extrair informações e recomendações baseadas nas mesmas

2 Informática na saúde

Nos anos 1960, 1970 e 1980, os sistemas de informação na saúde eram, em sua maioria, locais, isto é, voltados para departamentos e não para a instituição como um todo, além de serem utilizados principalmente para auxílio dos profissionais. Nos anos posteriores, os sistemas passaram a ser mais globais, considerando o processamento de informação da instituição, e também de regiões de saúde, e passaram a adotar abordagens mais centradas no paciente, utilizando estas informações principalmente para cuidados com o mesmo e para fins administrativos. Mais tarde, estas informações estenderam a possibilidade de uso dos dados, fazendo com que eles passassem a ser usados também para o planejamento de assistência médica e pesquisas clínicas (HAUX, 2006).

Outra mudança que ocorreu a partir dos anos 1990, foi a mudança no foco dos problemas, que passou de problemas técnicos, para problemas organizacionais, questões sociais e aspectos da gestão de mudanças. Houve também um maior reconhecimento da necessidade do gerenciamento e do planejamento dos sistemas de informação, além da inclusão de novos tipos de dados e tecnologias, ampliando as possibilidades de organização do sistema de saúde (HAUX, 2006).

Nos dias de hoje, os sistemas modernos de saúde utilizam dados de forma intensiva, uma vez que vários atores e entidades, como hospitais, clínicas, médicos, enfermeiros e muitos outros, necessitam da troca de grandes quantidades de informações instantaneamente. A Figura 1 representa a troca de dados entre diferentes atores em um exemplo de sistema de saúde.

Estes dados são principalmente de registros eletrônicos de saúde, conhecidos como EHR (*Eletronic Health Records*) e registros eletrônicos médicos, conhecidos como EMR (*Eletronic Medical Records*) (CASOLA et al., 2016), e, conforme as setas dos fluxos de comunicação da Figura 1, consistem principalmente de dados do paciente, dados médicos, dados de exames e dados das atividades do processo administrativo das instituições envolvidas no sistema de saúde.

Com as pesquisas na área de cidades inteligentes há a oportunidade de integração dos dados disponibilizados pela infraestrutura das cidades, como sensores, câmeras e relatórios meteorológicos, com os registros de saúde dos cidadãos, ampliando a diversidade de dados disponíveis para a criação de novas aplicações no contexto da saúde (SOLANAS et al., 2014).

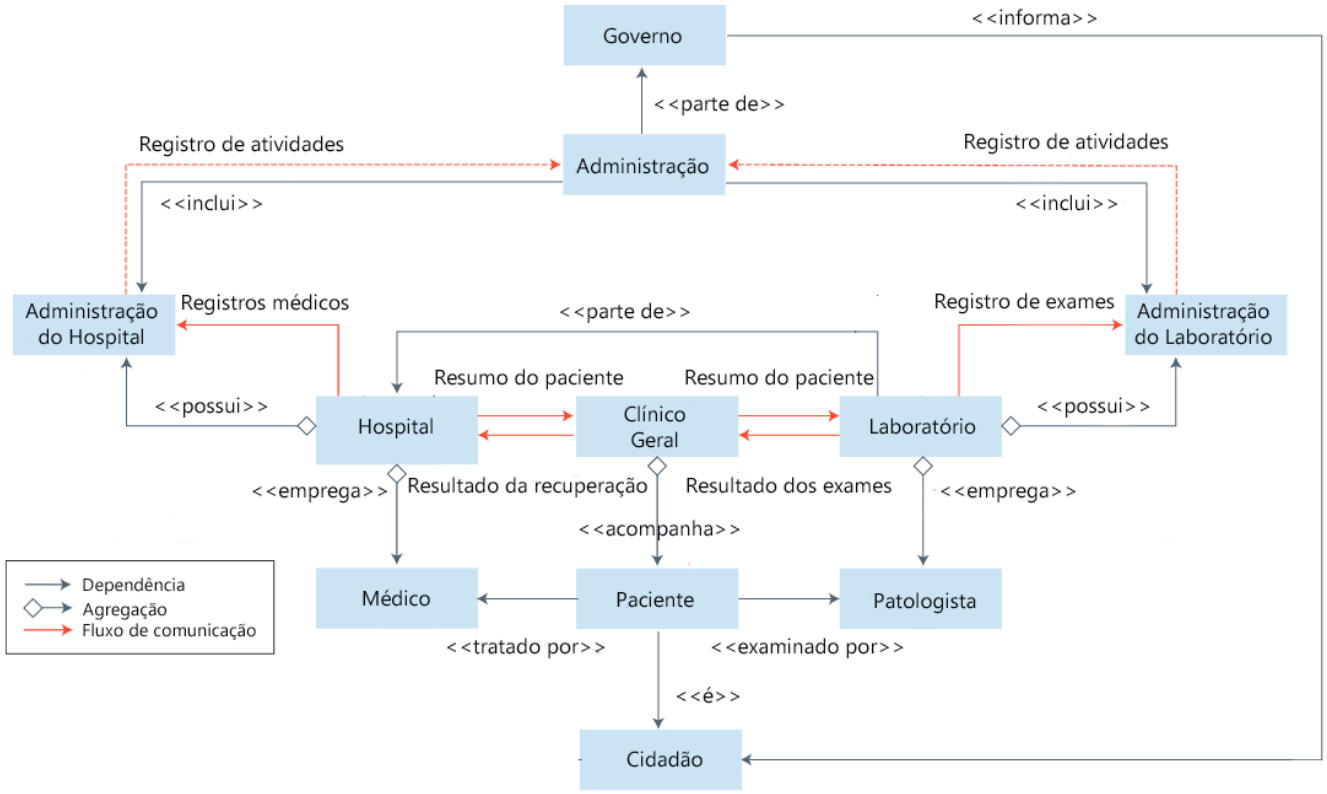

Figura 1 – Representação da troca de dados em um sistema de saúde (CASOLA et al., 2016) (tradução da autora).

2.1 *Open Source* na saúde

Apesar do seu aumento, a adoção de sistemas de informação na área da saúde possui alguns desafios, sendo o custo, a complexidade e a interoperabilidade alguns deles.

O custo é um desafio uma vez que a compra e os custos operacionais de sistemas de informação de saúde são muito altos. Outro desafio é a complexidade, que ocorre pois a transição para o sistema pode reduzir a produtividade nos períodos iniciais da adoção, caso o sistema seja muito complexo. Além disso, a necessidade de customizar o sistema para a realidade da instituição pode aumentar esta complexidade. Por fim, como o sistema de saúde envolve diferentes instituições, pacientes e profissionais, o dado precisa ser portável e compatível com diferentes sistemas, o que nem sempre ocorre, uma vez que as empresas privadas não querem que existam incentivos para que seus clientes migrem para outros serviços, fazendo com que os sistemas tenham baixa interoperabilidade (SAFADI et al., 2015), (YELLOWLEES et al., 2008).

Softwares open-source possuem seu código-fonte disponível para que qualquer pessoa revise, critique, modifique e redistribua (YELLOWLEES et al., 2008). Estas características fazem com que a adoção deste tipo de *software* na área da saúde se mostre benéfica em diferentes aspectos. O principal deles é o custo.

Em *softwares open-source*, o custo com a aquisição, manutenção e instalação do sistema é mais baixo do que em *softwares* proprietários. Isto ocorre tanto pela distribuição livre, economizando o gasto com a aquisição do *software*, quanto pela liberdade no suporte, uma vez que a própria instituição pode se responsabilizar pelo suporte. Em caso de projetos mais maduros, a própria comunidade se empenha em resolver problemas, como no caso do *TensorFlow*, uma biblioteca para aprendizado de máquina da Google, que se beneficia da disponibilização do código da biblioteca através da contribuição da comunidade encontrando erros, resolvendo-os, sugerindo novas funcionalidades ou até mesmo implementando-as (SAFADI et al., 2015).

Com a utilização de padrões e protocolos abertos, os *softwares open-source* também se mostram uma solução na melhoria da interoperabilidade. Uma vez que os sistemas sejam feitos sob os mesmos padrões, a comunicação entre eles é facilitada, além disso, o fácil acesso tanto ao código, quanto aos padrões por ele utilizados, permitem mais flexibilidade na mudança do sistema e na sua customização para a realidade da instituição (SAFADI et al., 2015), (REYNOLDS; WYATT, 2011).

Apesar dos *softwares open-source* ainda não terem ganho ampla aceitação na área da saúde, há alguns casos de sucesso, como o VistA (*Veterans Health Information Systems and Technology Architecture*) que possui versões em uso tanto nos Estados Unidos, quanto em outros países, como México; e o OpenMRS (*Open Medical Record System*, que segundo Alsaffar et al. (2017), foi um dos projetos *open-source* de saúde mais baixados em 2014 (YELLOWLEES et al., 2008).

Com isto, os modelos propostos neste trabalho serão *open-source*, a fim de diminuir custos e complexidade, além de permitir a comunicação com outros sistemas e a adaptação para diferentes contextos na área da saúde.

2.2 Trabalhos Relacionados

Durante a pesquisa para construção da base teórica também buscou-se alguns trabalhos relacionados à aplicação do BI e do *data science* na área de saúde, principalmente no contexto brasileiro e com enfoque na gestão. Três deles estão descritos a seguir.

Santos (2011) propõe um ambiente de BI para a gestão de informação em saúde na Secretaria Municipal de Saúde de Belo Horizonte. Foi construído um sistema usando principalmente relatórios e OLAP, contemplando 21 indicadores relacionados aos principais protocolos assistenciais da Secretaria de Saúde.

Ferraz (2009) propõe uma solução de BI para a Secretaria da Saúde do estado de Pernambuco. Através de uma modelagem de dados multidimensional, o OLAP foi utilizado na produção de gráficos e relatórios, contemplando principalmente informações

sobre os atendimentos e suas relações: paciente atendido, tipo de ocorrência, forma de transporte do paciente até o hospital, dentre outras informações.

Costa et al. (2008) propõe um ambiente de BI para a Saúde Pública na Secretaria Municipal de Saúde da Cidade de São Paulo, através também da modelagem multidimensional e OLAP, explorando os seguintes assuntos: agendamentos, atendimentos, equipes de saúde, estabelecimentos, fila de espera, pacientes, procedimentos, profissionais, regulação, vacinação e vagas. Os resultados foram reportados principalmente através de relatórios e gráficos.

A principal diferença entre os trabalhos encontrados e o trabalho proposto, é que neste trabalho propõe-se também a análise da utilização do aprendizado de máquina como técnica de análise dos dados. A utilização do aprendizado de máquina na saúde é presente utilizando dados mais clínicos, como em Gonçalves, Santos e Cruz (2010), onde é proposto um sistema de BI para a análise da qualidade de vida pré e pós-operatória dos pacientes, utilizando árvores de decisão, uma técnica de aprendizado de máquina.

Apesar da utilização do aprendizado de máquina em contextos mais gerenciais não ser tão presente no Brasil, em outros países é possível encontrar diferentes aplicações destas técnicas com este propósito. O *Group Health Cooperative* utiliza o aprendizado de máquina para proporcionar melhores serviços de saúde a custos mais baixos, um exemplo é a estratificação de seus pacientes por características demográficas e condições médicas, a fim de determinar quais grupos usam a maioria dos recursos e desenvolver programas para ajudar a educar essas populações, e para prevenir ou gerenciar suas condições. No *Seton Medical Center*, o aprendizado de máquina é usado para diminuir o tempo de permanência do paciente, evitar complicações clínicas, desenvolver melhores práticas, melhorar os resultados do paciente e fornecer informações aos médicos. E, no *Sierra Health Services*, é utilizado para identificar áreas para melhorias de qualidade, incluindo diretrizes de tratamento, grupos de gerenciamento de doenças e gerenciamento de custos (KOH; TAN et al., 2011).

3 *Data Science*

A grande quantidade de dados sendo produzidos de forma muito rápida não acontece apenas na área da saúde. Devido ao aumento da capacidade de armazenamento e processamento de dados e do uso da tecnologia na sociedade, há uma grande quantidade de dados que carregam o potencial de serem transformados em informações e/ou conhecimentos úteis em diversas áreas (SONG; ZHU, 2016), (PORTO; ZIVIANI, 2014), (AALST, 2014). Isto pode ser feito através de diferentes métodos, como o processamento estatístico, processamento de sinais, inteligência artificial, BI, *data science*, entre outros, sendo estes dois últimos o foco deste trabalho.

Com a necessidade do estudo da extração de conhecimento a partir de dados, surgiu o chamado *Data Science*, ou Ciência dos Dados (DHAR, 2013, p. 64). O *Data Science* conecta diferentes áreas, como banco de dados, computação, estatística, e também áreas não-matemáticas como comunicação e raciocínio ético, de forma a realizar adequadamente a coleta, preparação, análise, visualização, gerenciamento e preservação dos dados (STANTON, 2013).

Para Song e Zhu (2016), o *data science* deve incorporar quatro fatores: infraestrutura, ciclo de vida de análise dos dados, habilidades de gerenciamento de dados e disciplinas comportamentais. O ciclo de vida de análise dos dados refere-se às etapas necessárias para que a análise dos dados seja realizada, incluindo a análise do negócio, compreensão, preparação e integração dos dados, modelagem, avaliação, implantação e monitoramento; as habilidades de gerenciamento referem-se ao conhecimento relacionado à modelagem de dados; e as disciplinas comportamentais, referem-se às habilidades relacionadas à pessoas e negócios, como pensamento crítico, perguntas criativas e a comunicação com especialistas no domínio que não tenham muito conhecimento técnico. A Figura 2 descreve um modelo de ciclo de vida para projetos de *data science*.

O aumento da capacidade das organizações na obtenção de dados sobre seus processos, levou à crescente aplicação do *data science* na área dos negócios, de forma a melhorar a tomada de decisões nas organizações (PROVOST; FAWCETT, 2013). De forma semelhante, o *business intelligence* também visa a obtenção de conhecimentos úteis na tomada de decisões a partir dos dados. O Diagrama de Venn ilustrado na Figura 3 ilustra as intersecções entre estas duas áreas. Apesar dos objetivos semelhantes, estes dois conceitos se diferem na abordagem de suas análises, enquanto o foco do BI é na análise descritiva, o foco do *data science* é na análise preditiva e prescritiva (PAIXÃO; SILVA; TANAKA, 2015).

A análise descritiva busca, através dos dados, responder questões como “o que

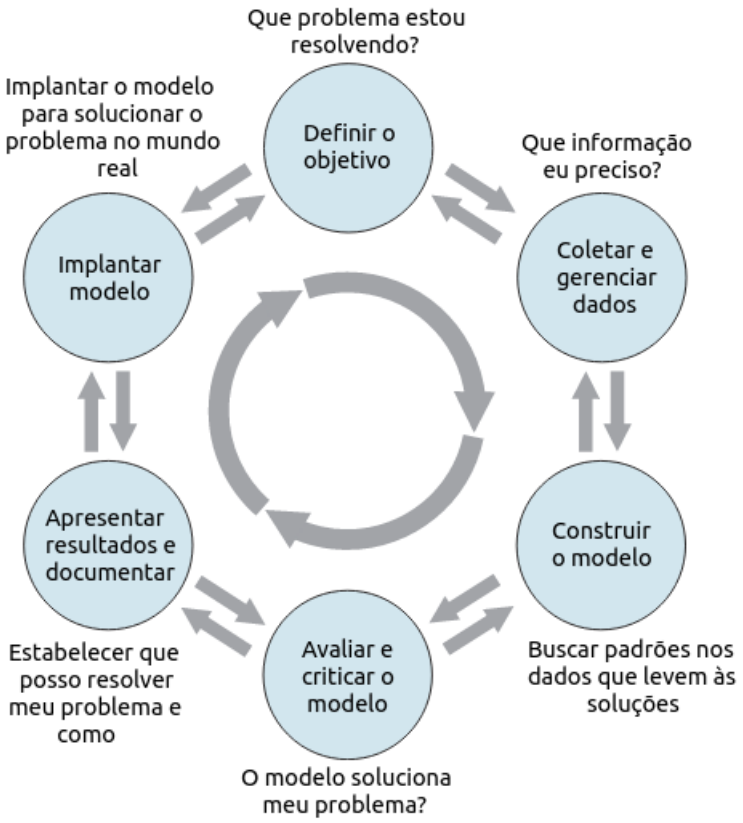

Figura 2 – Ciclo de vida de um projeto de *data science* (ZUMEL; MOUNT; PORZAK, 2014) (tradução da autora).

aconteceu?" e/ou "o que está acontecendo?", resultando principalmente na identificação de oportunidades e problemas de negócios. A análise preditiva busca responder questões como "o que acontecerá" e/ou "por que acontecerá?", utilizando das descobertas de padrões explicativos e preditivos (tendências, associações, afinidades, etc.) que representam as relações entre os dados. Por fim, a análise prescritiva visa a determinação de um conjunto de cursos ou ações alternativas, que, dado um conjunto de objetivos, requisitos e restrições, auxiliem na melhoria do desempenho do negócio (DELEN; DEMIRKAN, 2013). A Figura 4 resume estes três tipos de análise.

Dessa forma, estes dois conceitos, *business intelligence* e *data science*, se mostram complementares, uma vez que explorando seus diferentes tipos de análises é possível realizar uma análise mais completa da organização.

3.1 *Business Intelligence*

O termo *business intelligence* foi descrito por Howard Dresner, do Gartner Group, em 1989, como "um conjunto de conceitos e métodos para melhorar tomadas de decisões de negócios usando sistemas de suporte baseado em fatos" (ROUHANI; ASGARI;

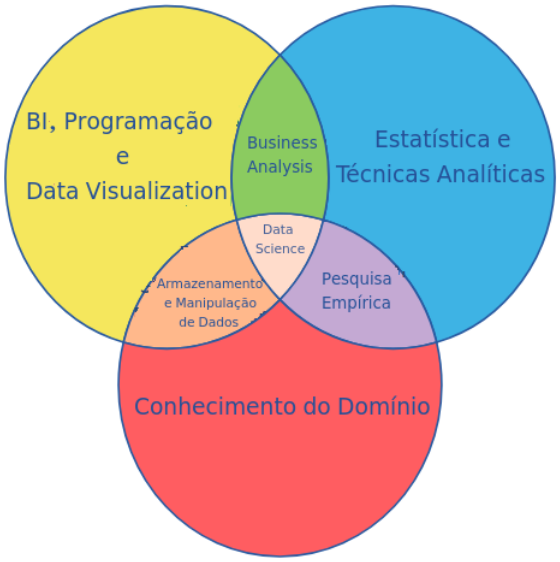

Figura 3 – Diagrama de Venn demostrando as intersecções entre o BI e o *data science* (AYANKOYA; CALITZ; GREYLING, 2014) (tradução da autor).

MIRHOSSEINI, 2012, p. 63). Apesar do termo ser relativamente novo, desde os anos 1970 aplicações de suporte à decisão têm sido utilizadas, entretanto, com a evolução da tecnologia e das necessidades das organizações, novas aplicações deste tipo foram surgindo, ampliando o domínio do suporte à decisão (WATSON; WIXOM, 2007).

Para Rouhani, Asgari e Mirhosseini (2012), o *business intelligence* é o processo de utilização e análise da informação para apoiar a tomada de decisões. Este processo utiliza de tecnologias que fornecem recursos para reunir, acessar e analisar dados do processo da organização, a fim de tornar a tomada de decisões melhor e mais rápida. Segundo Watson e Wixom (2007), o processo de *business intelligence* inclui duas atividades primárias: a entrada e a saída de dados. Na primeira delas, os dados são extraídos de diferentes fontes e transformados de modo a se tornarem significativos para o suporte à decisão. Na segunda atividade, já com os dados transformados, são realizadas análises através de *queries*, mineração de dados, relatórios, dentre outras técnicas que permitem a extração de informações úteis para responder às questões necessárias para a tomada de decisão. A Figura 5 ilustra estas duas atividades.

3.1.1 Entrada de dados

Assim como em outros tipos de organizações, na área da saúde os dados podem vir de diferentes fontes. Estas fontes podem ser internas, como registros de saúde eletrônicos, ou externas, como fontes governamentais, laboratórios, farmácias, entre outros. (RAGHUPATHI; RAGHUPATHI, 2014). Estas diferenças podem ocasionar dados com variações em sua qualidade, utilizando diferentes representações, códigos, formatos, além de residirem em diferentes locais. Devido a isso, para que as análises sejam feitas, é necessário que

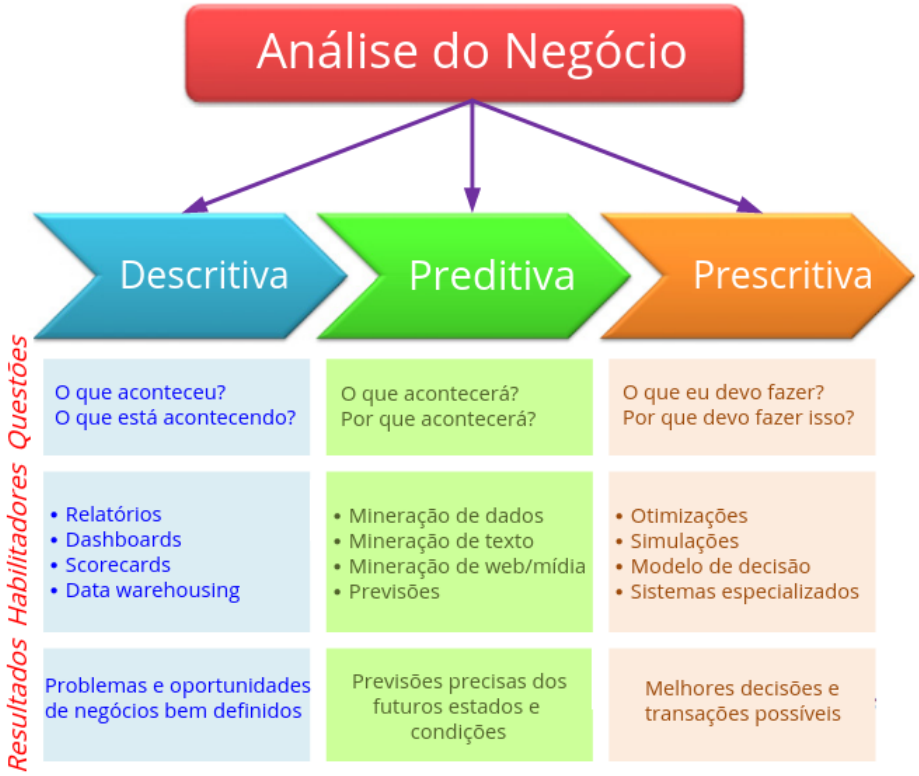

Figura 4 – Descrição dos tipos de análise (DELEN; DEMIRKAN, 2013) (tradução da autora).

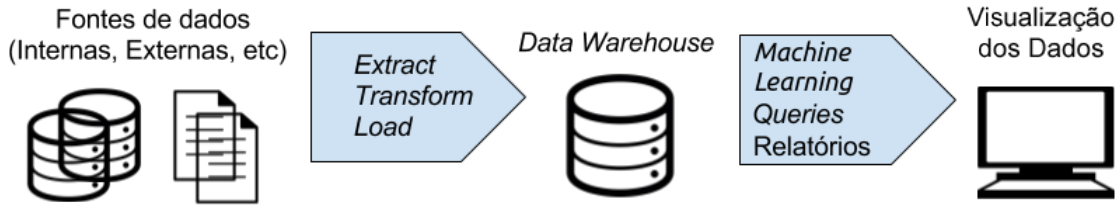

Figura 5 – Processo de *business intelligence*.

os dados sejam antes integrados, limpos, padronizados e disponibilizados para processamento (CHAUDHURI; DAYAL; NARASAYYA, 2011), (RAGHUPATHI; RAGHUPATHI, 2014).

Segundo Chaudhuri, Dayal e Narasayya (2011), as tecnologias responsáveis pela preparação dos dados são chamadas de ferramentas ETL (*Extract-Transform-Load*). Um sistema ETL apropriado, tira os dados das fontes, implementa padrões de qualidade e consistência, de modo que eles entrem em conformidade, e disponibiliza os dados, carregando-os em um único *data warehouse* (KIMBALL; CASERTA, 2011), conforme Figura 6. Há abordagens, como a virtualização de dados, em que os dados não são fisicamente movidos

das fontes, mas sim integrados a um único banco lógico (LANS, 2012). Arquiteturas orientadas a serviços permitem que o dado permaneça "bruto" e serviços sejam utilizados para obtê-lo, processá-lo e disponibilizá-lo (RAGHUPATHI; RAGHUPATHI, 2014). Uma vez que o dado é adequadamente preparado e disponibilizado, é possível realizar a atividade de saída de dados, onde ocorre a análise.

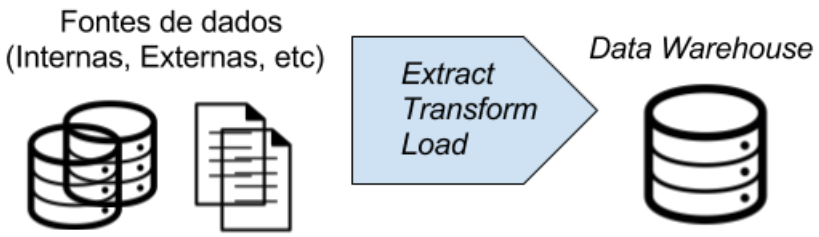

Figura 6 – Processo de entrada de dados.

3.1.2 Saída de dados

Segundo Watson e Wixom (2007), a obtenção dos dados proporciona um valor limitado à organização. Para obter total aproveitamento do *data warehouse*, é preciso que os dados sejam acessados por usuários e aplicações, e utilizados para a tomada de decisões, através de relatórios, *queries*, OLAP (*Online Analytic Processing*), aprendizado de máquina, entre outras técnicas (CHAUDHURI; DAYAL; NARASAYYA, 2011), conforme Figura 7.

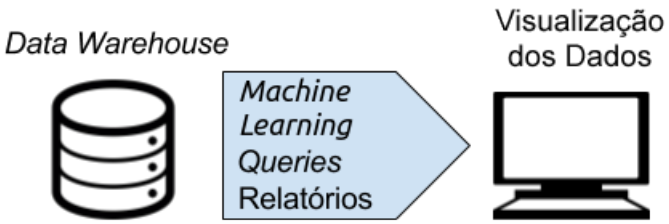

Figura 7 – Processo de saída de dados.

É possível a definição e renderização de relatórios que agregam informações sobre o contexto, como por exemplo, relatório do total de internações por idade, região, entre outros. Relatórios que utilizam de banco de dados relacionais precisam executar complexas *queries* SQL em grandes volumes de dados para obter estas informações, fazendo com que diferentes técnicas e estruturas de dados fossem desenvolvidas para viabilizar

estas ações, como por exemplo, a otimização de *queries*, *materialized views* e a indexação (CHAUDHURI; DAYAL; NARASAYYA, 2011). Outras técnicas como o OLAP e o aprendizado de máquina, são abordadas nos tópicos seguintes.

3.1.2.1 OLAP

O OLAP utiliza a modelagem dimensional dos dados. Nela os objetos de análise são medidas numéricas onde cada uma delas é associada à uma dimensão. Estas dimensões representam entidades do contexto, as entidades relacionadas à uma internação podem ser a data, a cidade e o paciente, por exemplo. Os atributos de uma dimensão podem ser relacionados através de uma hierarquia de relacionamentos, conforme a Figura 8 (CHAUDHURI; DAYAL; NARASAYYA, 2011), (CHAUDHURI; DAYAL; GANTI, 2001).

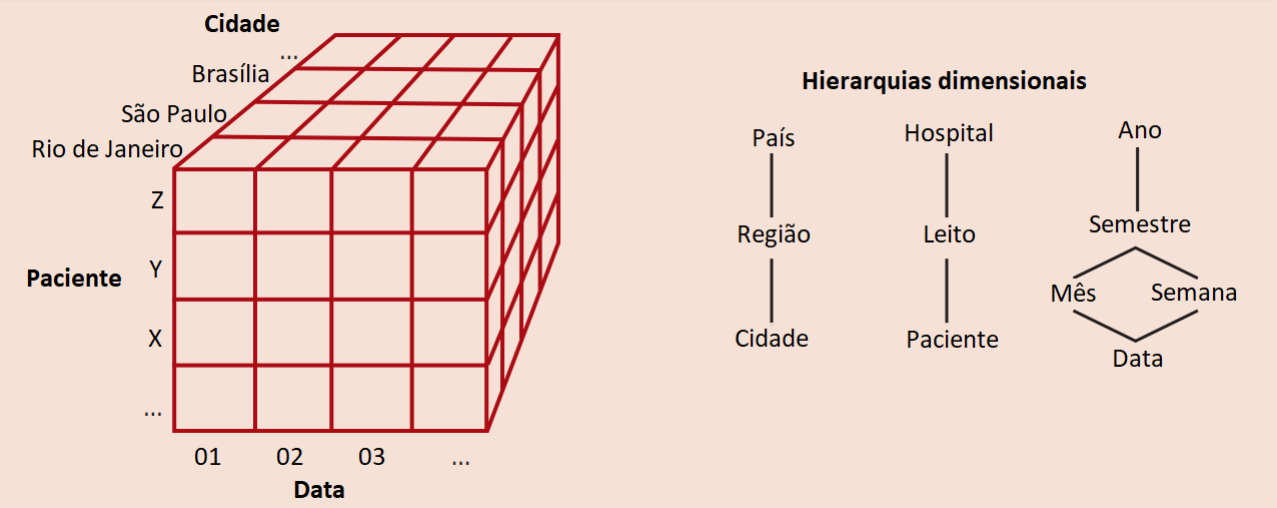

Figura 8 – Visão multidimensional dos dados (CHAUDHURI; DAYAL; NARASAYYA, 2011), (CHAUDHURI; DAYAL; GANTI, 2001) (adaptado pela autora).

Através do OLAP é possível fazer as operações de *slice*, *dice*, *drill-down*, *roll-up* e *pivoting*, nos cubos de dados. O *pivoting* permite a rotação do cubo, permitindo que ele seja analisado de diferentes ângulos. *Roll-up* e *drill-down* permitem, respectivamente, o aumento e a redução do nível de granularidade na hierarquia de uma dimensão. O *slice* é a extração de dados de uma única dimensão do cubo, enquanto o *dice* é a extração de um subcubo, através da extração de uma ou mais dimensões do cubo (HAN, 1997).

O OLAP pode ser implementado usando uma *engine* de armazenamento multidimensional (MOLAP), banco de dados relacionais (ROLAP), ou uma abordagem híbrida (HOLAP). *Engines* de armazenamento multidimensional utilizam uma abstração em *array* onde grandes cubos de dados são pré-computados. Na abordagem ROLAP, o modelo multidimensional é convertido em relacionamentos e *queries* SQL, e para isso são usados os chamados Esquema Estrela e Esquema Floco de Neve. No Esquema Estrela, o banco de dados consiste de uma única Tabela Fato e uma tabela para cada dimensão, sendo que

cada linha na Tabela Fato é uma *Foreign Key* para cada Tabela Dimensão, e cada Tabela Dimensão possui colunas que correspondem aos atributos da dimensão correspondente. Para o suporte à hierarquia de atributos, o Esquema Estrela foi refinado, resultando no Esquema Floco de Neve, onde as hierarquias dimensionais são representadas pela normalização das Tabelas Dimensão (CHAUDHURI; DAYAL; NARASAYYA, 2011).

3.1.2.2 Aprendizado de máquina

"O aprendizado de máquina é o estudo de métodos computacionais para automatizar o processo de aquisição de conhecimento a partir de exemplos" (BOSE; MAHAPATRA, 2001, p. 212). A Figura 9 descreve um processo pelo qual algoritmos de aprendizado de máquina podem ser selecionados, aplicados e avaliados para um problema.

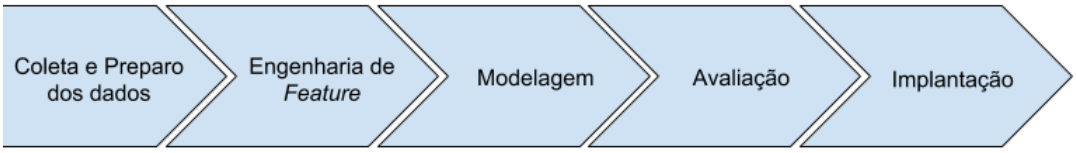

Figura 9 – Processo do aprendizado de máquina.

Para utilização de algoritmos de aprendizado de máquina, é necessário, inicialmente, que se tenha acesso ao dado e que ele esteja limpo. Com acesso ao dado é possível então identificar e construir *features* mais úteis, e usá-las na modelagem, onde são escolhidos os algoritmos mais apropriados para o problema e é construído um modelo com base nestes algoritmos e nos dados. O modelo então é avaliado, a fim de identificar sua capacidade de processar novos dados corretamente. E por fim, com uma avaliação aceitável, o modelo pode ser implantado para a resolução do problema no mundo real (MARSLAND, 2015), (WITTEN et al., 2016).

3.1.2.2.1 Engenharia de *feature*

Uma *feature*, também chamada de atributo, é a descrição de alguma característica ou aspecto do exemplo. Elas podem ser divididas em dois tipos: nominais e contínuas. *Features* nominais descrevem valores que não possuem uma ordem entre si, como por exemplo, o sexo de uma pessoa. Já *features* contínuas descrevem valores que possuem uma ordem linear, como por exemplo, a altura de uma pessoa (MONARD; BARANAUSKAS, 2003).

Muitas vezes assume-se que todas as *features*, isto é, todas as características do exemplo são relevantes, entretanto, podem existir atributos que não são diretamente relevantes, ou até mesmo irrelevantes. Por exemplo, para diagnosticar se uma pessoa está ou não com gripe, há atributos pouco relevantes, como a cor do olho da pessoa, e há atributos muito relevantes, como a temperatura da pessoa. Além da relevância, também

é importante analisar a distribuição das *features*, uma vez que distribuições homogêneas, por exemplo, não terão tanto impacto no resultado, e logo, poderão ser excluídas. É de suma importância para o sucesso do aprendizado, que as *features* mais significativas para o contexto sejam utilizadas (MONARD; BARANAUSKAS, 2003), (BLUM; LANGLEY, 1997).

Este processo de seleção de quais *features* serão utilizadas é chamado de *Feature Engineering*. Além de remover *features* irrelevantes e redundantes, o processo também inclui a combinação de *features*, uma vez que *features* que pareçam irrelevantes isoladas, podem se mostrar relevantes quando combinadas (DOMINGOS, 2012), (MAGLOGIANNIS, 2007).

3.1.2.2.2 Modelagem

Na modelagem, os algoritmos de aprendizado de máquina são escolhidos de acordo com os dados e com o problema, e um modelo é construído com o objetivo de fazer uma descrição estrutural a informação explícita no dado (WITTEN et al., 2016), (MARSLAND, 2015). Estes algoritmos podem ser divididos em duas categorias: aprendizado supervisionado e aprendizado não-supervisionado. A Figura 10 descreve a relação destas duas categorias com os tipos de problemas para os quais cada uma é apropriada.

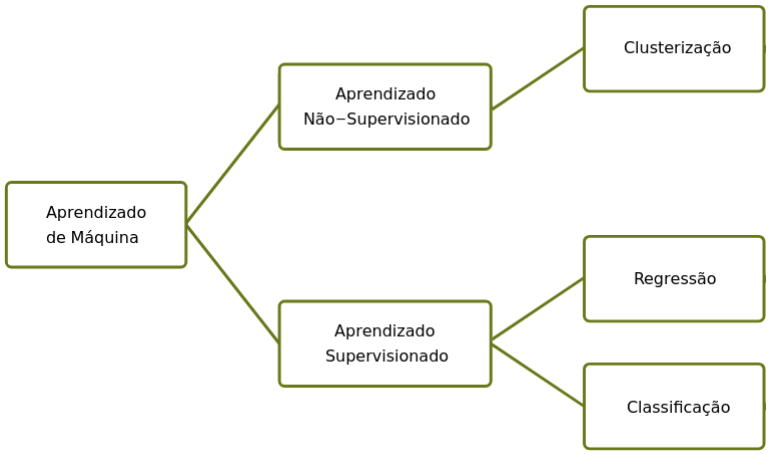

Figura 10 – Tipos de aprendizado de máquina.

Aprendizado Supervisionado

No aprendizado supervisionado, o algoritmo recebe exemplos de treinamento especificando entradas e a saída correta para aquelas entradas. O algoritmo processa os exemplos, extraindo o aprendizado sobre eles. Quando novas entradas forem submetidas ao algoritmo, ele deve ser capaz de produzir as saídas corretas (LORENA; CARVALHO, 2007). A Figura 11 descreve este processo.

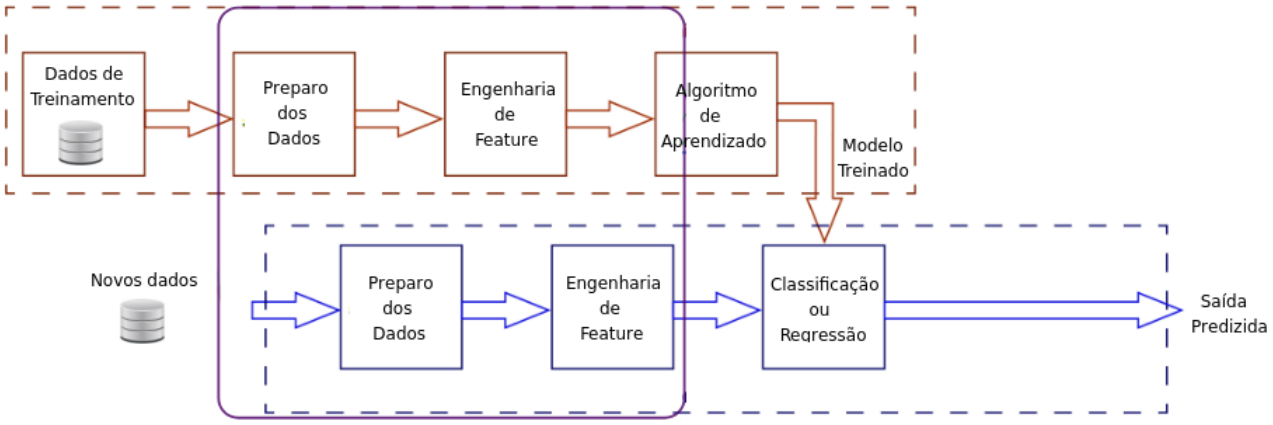

Figura 11 – Processo de aprendizado supervisionado.

Para um exemplo prático, supõe-se um conjunto de dados onde há o registro com informações de vários pacientes, onde cada registro possui os sintomas do paciente e a classificação do atendimento que ele necessita. O algoritmo irá receber este conjunto de dados, de forma que os sintomas serão as entradas, e a classificação corresponde será a saída. Dessa forma, após processar e adquirir o conhecimento a partir destes exemplos, quando sintomas de novos pacientes forem submetidos ao algoritmo, ele deve saber identificar a qual classificação de atendimento o paciente necessita.

A saída desejada, isto é, a meta a qual se deseja aprender e fazer previsões a respeito é chamada de classe ou rótulo. No exemplo anterior, deseja-se fazer previsões a cerca da classificação do atendimento do paciente, logo, esta é a classe. Os dados de treinamento são chamados de dados rotulados, uma vez que possuem a saída especificada. Quando os rótulos possuem valores discretos, o problema de determinar corretamente a classe de novos exemplos ainda não rotulados é chamado de **classificação**, já quando possuem valores contínuos, tem-se um problema de **regressão** (MONARD; BARANAUSKAS, 2003). Alguns dos principais algoritmos utilizados para classificação são as árvores de decisão, redes neurais e o KNN (*K-Nearest Neighbor*), já para regressão, são usadas principalmente a regressão linear simples e a regressão linear múltipla (LAROSE, 2014).

Para evitar resultados tendenciosos, no aprendizado supervisionado costuma-se realizar a partição dos dados em subconjuntos, sendo eles de **treinamento, validação e teste**. O algoritmo é inicialmente submetido à um subconjunto de treinamento, enquanto os demais subconjuntos são utilizados para uma posterior confirmação da análise inicial, sendo o de validação para avaliar o aprendizado do algoritmo, e o de teste para produção dos resultados finais (TAVARES; LOPES; LIMA, 2007). É importante que os subconjuntos sejam escolhidos de forma a manter a representatividade da distribuição dos exemplos no mundo real, evitando *overfitting* e *underfitting* (MARSLAND, 2015).

É possível que o subconjunto de treinamento escolhido induza hipóteses que me-

lhorem o desempenho do algoritmo no conjunto de treinamento, enquanto pioram o desempenho nos demais subconjuntos. Neste caso, o erro (ou outra medida) nos conjuntos de validação evidencia um desempenho ruim na hipótese. Este fenômeno é chamado de *overfitting*, e ocorre quando a hipótese se ajusta em excesso ao subconjunto de treinamento (MONARD; BARANAUSKAS, 2003).

É possível também que sejam escolhidos subconjuntos de forma que ocorra uma melhora de desempenho muito pequena no conjunto de treinamento, assim como em um conjunto de teste. Este fenômeno é chamado de *underfitting*, e ocorre quando a hipótese ajusta-se muito pouco ao conjunto de treinamento (MONARD; BARANAUSKAS, 2003).

Aprendizado Não-Supervisionado

No aprendizado não-supervisionado, os exemplos não são rotulados, dessa forma, não há uma classe a qual se deseja prever. Os algoritmos recebem os exemplos não rotulados e tentam agrupá-los conforme sua similaridade. Estes agrupamentos são chamados de *clusters*, por isto, este tipo de problema recebe o nome de clusterização. Normalmente, depois do agrupamento é necessária uma análise para identificar o que cada grupo significa naquele contexto (MONARD; BARANAUSKAS, 2003). Alguns dos principais algoritmos de clusterização são o *clustering* hierárquico, *k-means* e as redes neurais.

Um exemplo prático do aprendizado não-supervisionado é a categorização de *e-mails* por assunto. Dado um conjunto de *e-mails*, o algoritmo os recebe como exemplo e tenta agrupá-los analisando as similaridades em seu conteúdo. Diferentemente do aprendizado supervisionado, o algoritmo não é previamente treinado com *e-mails* com assuntos já definidos, mas sim, cria grupos com os exemplos que aparentam ter alguma similaridade, por isso pode ser necessária uma análise para identificar os significados dos grupos, nesse caso, para identificar a qual assunto o grupo se refere.

3.1.2.2.3 Avaliação

Não existe apenas um algoritmo de aprendizado de máquina que apresente o melhor desempenho para todo tipo de problema. Cada problema tem suas peculiaridades, e portanto, deve ser analisado o poder e as limitações do algoritmo naquele contexto. Para isto, existem diferentes métodos de avaliação (MONARD; BARANAUSKAS, 2003).

Estes métodos fornecem maneiras de avaliar o desempenho de um sistema de aprendizagem, permitindo identificar se os resultados realmente possuem valor preditivo ou se refletem falsas regularidades. Se a avaliação refletir que o modelo é ruim, pode ser necessário reconsiderar o algoritmo utilizado e/ou até mesmo as *features*, caso a avaliação reflita que o modelo é suficientemente bom, o próximo passo é implementá-lo na prática (WITTEN et al., 2016).

4 Metodologia

Este capítulo tem o objetivo de apresentar as metodologias (seção 4.1) e o planejamento (seção 4.2) que serão utilizados neste trabalho, a fim de atingir os objetivos da pesquisa.

4.1 Metodologias de Pesquisa

Segundo Prodanov e Freitas (2013), as formas clássicas de classificação de pesquisa são quanto à natureza da pesquisa, quanto aos seus objetivos e quanto aos seus procedimentos, conforme Figura 12.

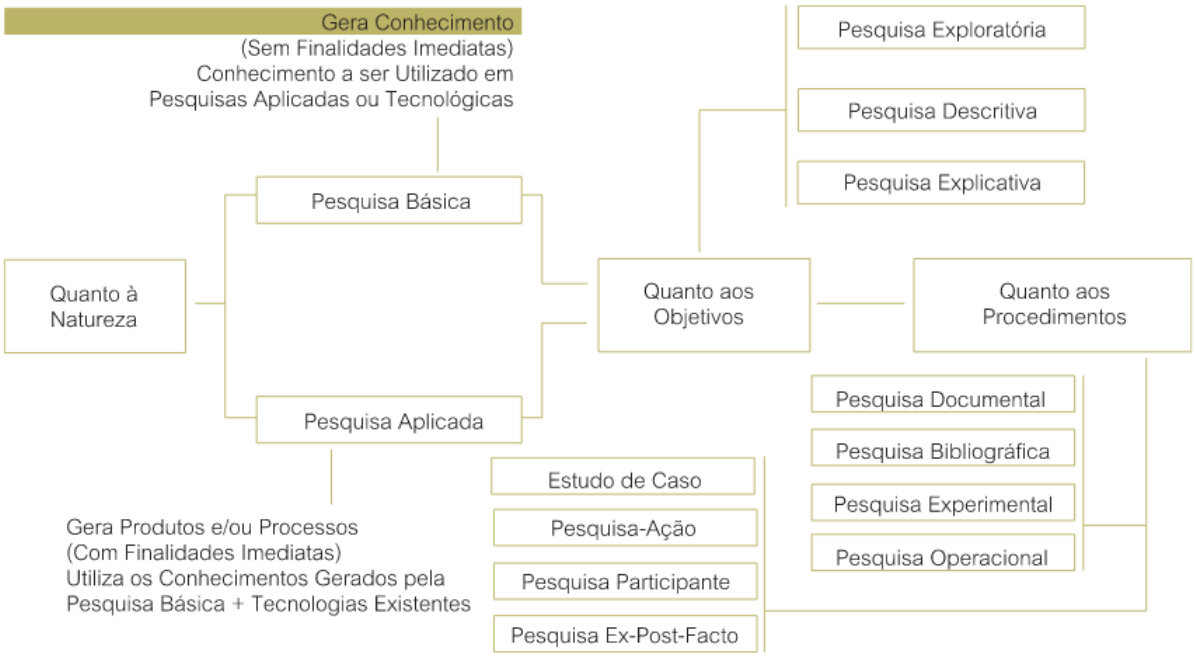

Figura 12 – Classificações de pesquisa científica (PRODANOV; FREITAS, 2013).

Quanto à sua natureza, uma pesquisa pode ser aplicada ou básica. A pesquisa básica visa gerar novos conhecimentos úteis para a ciência, porém sem a previsão de uma aplicação prática, já a pesquisa aplicada visa gerar conhecimentos para aplicação prática, solucionando problemas específicos (PRODANOV; FREITAS, 2013).

Quanto aos objetivos, a pesquisa pode ser exploratória, descritiva ou explicativa. A pesquisa exploratória visa proporcionar mais informações sobre o assunto estudado, permitindo a construção de hipóteses; a pesquisa descritiva busca registrar e descrever os fatos observados sem interferência do pesquisador; e a pesquisa explicativa busca explicar

o porquê das coisas e suas causas, identificando fatores que contribuem para determinados fenômenos (PRODANOV; FREITAS, 2013).

Quanto aos procedimentos existem dois grandes grupos: os que utilizam fontes de "papel" (pesquisa bibliográfica e pesquisa documental) e os que utilizam dados fornecidos por pessoas (pesquisa experimental, pesquisa ex-post-facto, o levantamento, o estudo de caso, a pesquisa-ação e a pesquisa participante). Destes dois grupos, destacam-se neste trabalho a pesquisa bibliográfica e o estudo de caso. A pesquisa bibliográfica é elaborada através de material já publicado, principalmente artigos, livros, revistas, jornais, dissertações, entre outros, e estudo de caso busca o estudo de um objeto através da aplicação prática de conhecimentos para a solução de problemas (PRODANOV; FREITAS, 2013).

Por fim, quanto à abordagem, uma pesquisa pode ser quantitativa ou qualitativa. Na pesquisa quantitativa, as opiniões e informações são traduzidas em números para então serem analisadas, já a pesquisa qualitativa considera que há uma subjetividade no sujeito que não pode ser traduzida em números (PRODANOV; FREITAS, 2013).

Diante das metodologias apresentadas, o presente trabalho se classifica como pesquisa aplicada, quanto à sua natureza; pesquisa exploratória, quanto aos seus objetivos; apresenta tantos aspectos quantitativos quanto qualitativos quanto à sua abordagem; e utiliza, como procedimentos, a pesquisa bibliográfica e o estudo de caso.

4.2 Planejamento da Pesquisa

Foi definido como caso de estudo a Secretaria de Saúde do Distrito Federal e a Secretaria de Saúde de São Paulo, uma atuando no fornecimento de pessoas com domínio do contexto, e a outra no fornecimento dos dados a serem utilizados no trabalho, respectivamente, a fim de contribuir para a resolução da seguinte questão de pesquisa:

Qual ou quais estratégias podem ser utilizadas no business intelligence de forma a melhor auxiliar no uso de dados para tomada de decisões na área da saúde?

A fim de responder à esta questão, as seguintes hipóteses foram formuladas:

- H1: Técnicas clássicas de *business intelligence*, como relatórios, *queries* e OLAP são úteis para análises descritivas.
- H2: O aprendizado de máquina complementa a análise de dados, possibilitando análises preditivas e prescritivas.

Com base nas hipóteses, nas metodologias definidas e nos ciclos de vida de projetos de BI e *data science*, foi elaborado o fluxo de trabalho ilustrado na Figura 13.

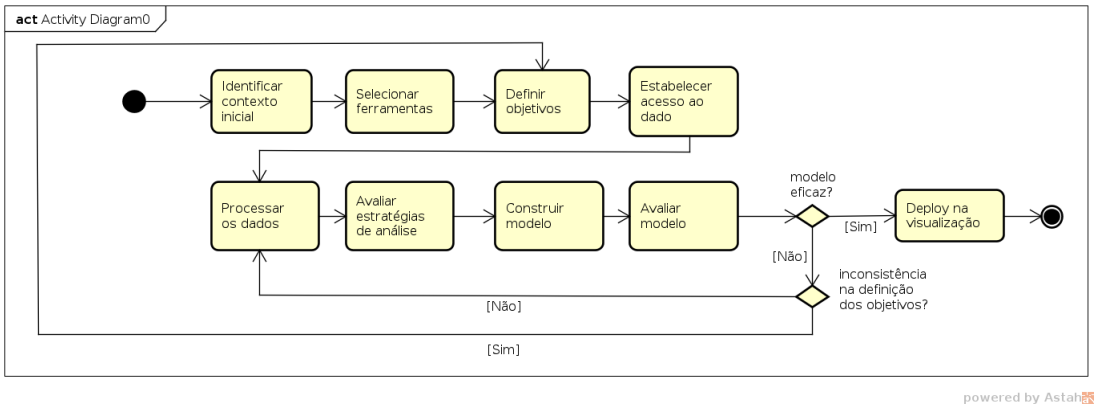

Figura 13 – Fluxo de atividades para elaboração dos modelos.

- **Identificar o contexto inicial** - esta atividade refere-se à identificação do estado atual do contexto, isto é, quais os problemas na solução atual, quais ferramentas e dados utilizadas na solução atual, quais as necessidades da organização, dentre outras informações.
- **Selecionar as ferramentas** - diante das informações sobre o estado atual do contexto, esta atividade diz respeito à seleção das ferramentas que serão utilizadas no BI, a fim de atingir as necessidades identificadas.
- **Definir objetivos** - com as ferramentas selecionadas, agora é possível utilizá-las para extração de informações úteis para a gestão, mas antes é necessário entender quais informações devem ser obtidas através do BI, quais dados são necessários para gerar estas informações e se estes dados estão disponíveis.
- **Estabelecer acesso adequado ao dado** - com os dados definidos e disponibilizados na atividade anterior, é possível agora, nesta atividade, integrá-los às ferramentas selecionadas.
- **Processar os dados** - com acesso ao dado é possível então prepará-lo, limpando-o e identificando novas *features* para serem usadas nos modelos, tarefas realizadas nesta atividade.
- **Avaliar estratégias de análise dos dados** - com os dados preparados é possível então analisar, para os problemas definidos, quais estratégias serão mais eficazes.
- **Construir modelo** - nesta atividade os dados são modelados de forma a resolver o problema utilizando a estratégia definida. É possível que essa atividade também inclua tarefas de preparação dos dados, uma vez que ela busca a melhor forma de representar e modelar o dado (ZUMEL; MOUNT; PORZAK, 2014).
- **Avaliar o modelo** - por fim, deve ser avaliado se o modelo realmente soluciona os problemas definidos, se não solucionar é necessário identificar o motivo, caso seja

uma falha na definição dos problemas, é necessário voltar para esta atividade, a fim de analisar os problemas e os dados novamente. Caso os problemas tenham sido bem definidos, é necessário voltar para a atividade de processamento dos dados, a fim de identificar dados faltantes, necessidade de dados serem mais limpos, novas *features* possíveis, entre outras soluções, reanalisando também, posteriormente, as escolhas das estratégias de análise.

- ***Deploy* na visualização** - Com o sucesso do modelo, é possível por fim promover a visualização dos resultados na ferramenta.

Segundo Larson e Chang (2016), o dinamismo requerido pelo BI faz com que uma abordagem ágil seja apropriada para estes projetos. Os princípios ágeis defendem principalmente os indivíduos e interações sobre processos e ferramentas, *software* funcional sobre documentação completa, colaboração com o cliente sobre negociação de contratos e resposta às mudanças sobre seguir um plano.

Desta forma, o fluxo de atividades da Figura 13 foi elaborado visando uma execução iterativa e incremental, garantindo principalmente a adaptação às mudanças, entrega contínua e comunicação com os envolvidos. As atividades serão executadas utilizando em *sprints*, um *time-box* de duas semanas no qual um incremento do produto é criado (SCHWABER; SUTHERLAND, 2016).

Todo trabalho será realizado como *software* livre, sob a licença GPL (*General Public License*) da GNU, versão 3. O acompanhamento do projeto poderá ser feito pelo GitHub.

4.2.1 Identificação do contexto

O projeto é voltado para o contexto da gestão hospitalar, uma vez que os dados obtidos para o projeto, até o momento, são os dados que compõem o prontuário médico. Desta forma, conforme a Figura 1 do Capítulo 2, este contexto irá lidar principalmente com as entidades: Hospital, Administração do Hospital e Paciente.

Em contato com as pessoas envolvidas no contexto, também foi possível identificar alguns problemas inicias, sendo eles:

- **Baixa interoperabilidade** - a dificuldade em acessar os dados do sistema proprietário fazem com que seja necessária a criação de vários outros sistemas para manipulação dos dados.

- **Difícil customização** - a limitação no acesso aos dados e ao código dos *softwares* proprietários dificulta também a customização das funcionalidades.

- **Sistemas obsoletos** - como o custo de atualização e manutenção dos sistemas proprietários é alto, eles se tornam obsoletos.

- **Ausência de sistemas de BI e/ou geração de relatórios** - as ferramentas atuais permitem o registro de muitos dados, entretanto há uma grande dificuldade em gerar relatórios e outras análises destes dados.

4.2.2 Seleção das ferramentas

Com base no contexto inicial e nas hipóteses elaboradas, foram identificados os critérios necessários para a seleção das ferramentas utilizadas no *business intelligence*. Foram consideradas tanto ferramentas completas de BI, quanto a utilização conjunta de ferramentas específicas para processamento e visualização dos dados.

4.2.2.1 Critérios

Foram analisados os critérios sugeridos por Brandão et al. (2016), Lapa, Bernardino e Figueiredo (2014) e Oestreich (2016). Estes critérios foram selecionados de acordo com a relevância para o contexto específico, resultando nos seguintes critérios finais: *open source*, performance, escalabilidade, integração com aprendizado de máquina, conectividade com as fontes, ETL, OLAP, customização, configuração/instalação e facilidade de uso.

- *Open-Source*

  Este critério foi considerado uma vez que em ferramentas *open-source*, como o código é disponibilizado, é possível fazer adaptações de forma que o sistema se adeque melhor ao contexto da organização, além disso, a comunidade pode contribuir com resoluções de erros e evoluções, e não há custos com a aquisição do sistema.

- Performance

  Este critério diz respeito à performance da ferramenta no processamento de dados. Neste contexto, uma boa performance é importante uma vez que na área da saúde há grandes volumes de dados e, além disso, as decisões podem ter grande impacto na vida de seres humanos.

- Escalabilidade

  Este critério diz respeito à capacidade da ferramenta em manter um bom processamento mesmo quando a quantidade de trabalho aumenta. Este critério é importante uma vez que a quantidade de dados pode aumentar, bem como as necessidades de processamento, e é importante que nestes casos o sistema continue em funcionamento.

- Integração com aprendizado de máquina

  Como uma das hipóteses é que o aprendizado de máquina tornará a análise mais completa, é importante que a ferramenta permita a integração com algoritmos de aprendizado de máquina.

- Conectividade com as fontes

  Na área da saúde os dados podem vir de diferentes fontes e em diferentes formatos (RAGHUPATHI; RAGHUPATHI, 2014), dessa forma, é importante que a ferramenta permita a integração com múltiplas e diferentes fontes de dados.

- ETL

  Este critério considera a possibilidade de executar passos de extração, transformação e carregamento dos dados na própria ferramenta.

- OLAP

  Este critério avalia a possibilidade de realizar análises multidimensionais nos dados através do OLAP.

- Customização

  Uma vez que um dos problemas identificados no contexto do caso de estudo foi a dificuldade de customização das ferramentas vigentes, é importante que a ferramenta seja customizável de forma a melhor atender as particularidades da organização.

- Configuração/Instalação

  O processo de BI demanda um grande tempo e esforço nas atividades de extração, transformação e carregamento dos dados (WATSON; WIXOM, 2007), dessa forma, é importante que a configuração e instalação da ferramenta seja prática, de forma a não prejudicar o tempo disponível para as outras etapas. Além disso, nem sempre os profissionais envolvidos possuem conhecimento a cerca da ferramenta, dessa forma, é importante que a configuração e a instalação sejam práticas para que estas pessoas também consigam utilizar a ferramenta em tempo hábil.

4.2.2.2 Ferramentas

Há tanto ferramentas específicas para diferentes etapas do BI, quanto ferramentas gerais que integram funcionalidades úteis em várias etapas. De acordo com os critérios estabelecidos e com o contexto do projeto, foram avaliadas ferramentas específicas, de forma a selecionar um conjunto que contemple todo o processo de BI, e também ferramentas completas, comparando-as ao cojunto selecionado de forma a chegar a um resultando final.

4.2.2.2.1 Ferramentas completas

Segundo o *benchmarking* realizado por Brandão et al. (2016), analisando ferramentas *open-source* de BI em ambientes de saúde, as ferramentas que mais se destacaram foram o Spago BI e o Pentaho. Estas ferramentas também se destacaram na comparação realizada por Lapa, Bernardino e Figueiredo (2014). Dessa forma, foi realizada uma análise destas duas ferramentas.

Pentaho

O Pentaho é um projeto desenvolvido pela *Pentaho Corporation*, o qual consiste em diversos produtos, sendo eles: *Pentaho BI Platform*, *Pentaho Reporting*, *Pentaho Analysis*, *Pentaho Data Integration*, *Community Edition Dashboard* e *Weka Pentaho Data Mining*. Através da integração destes produtos, o Pentaho permite a construção de relatórios, análise OLAP, suporte à ETL, construção de *dashboard* e análises preditivas (BRANDÃO et al., 2016). Um dos principais problemas do Pentaho é a dificuldade em sua utilização, ficando com baixo ranqueamento em diversos aspectos desta categoria na avaliação realizada por Oestreich (2016). A Figura 14 demonstra um exemplo de *dashboard* criada com o Pentaho.

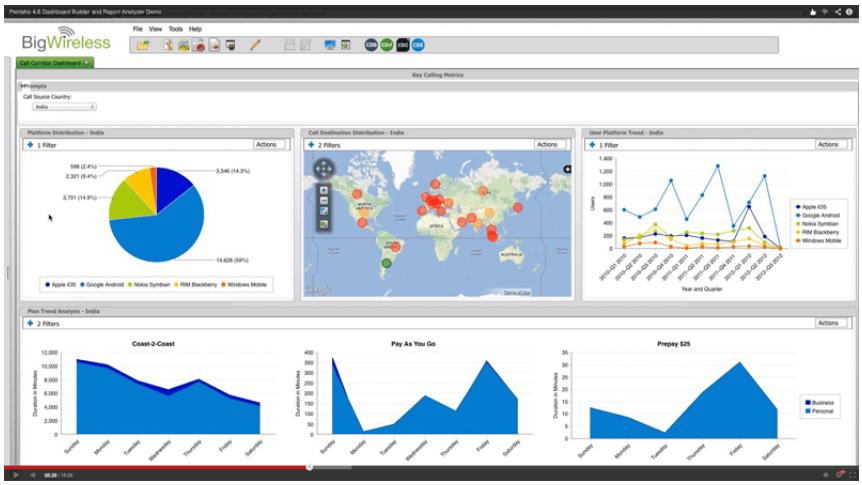

Figura 14 – Exemplo de *dashboard* criada com o Pentaho.

Fonte: <http://www.techtem.com.br/pentaho-software-bi/>. Acesso em 05 de março de 2018

Spago BI

O Spago BI também é uma ferramenta constituída por diversos módulos, sendo eles: *Spago BI Server*, *Spago BI Studio*, *Spago BI Meta*, *Spago BI SDK* e *Spago BI Applications*. O *Spago BI Server* é o módulo principal, constituindo o *core* da aplicação. O

Spago BI Studio é um ambiente de desenvolvimento baseado no Eclipse, ele permite a modificação dos documentos de análise, como relatórios, OLAP, *dashboards* e mineração de dados. O *Spago BI Meta* é responsável pela manipulação dos dados e processos de ETL. O *Spago BI SDK* é a ferramenta responsável pela integração dos serviços providos pelo servidor. E por fim, o *Spago BI Applications* é uma coleção de modelos analíticos desenvolvidos usando o Spago BI (BRANDÃO et al., 2016). A Figura 15 demonstra um exemplo de *dashboard* criada com o Spago BI.

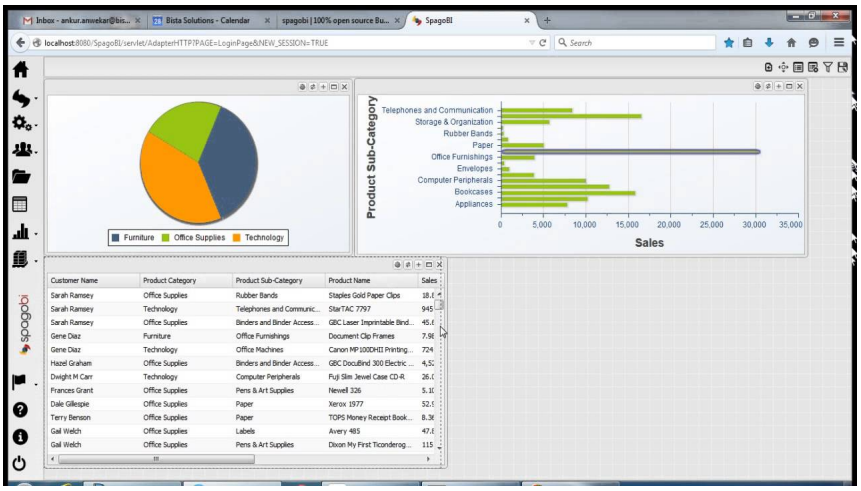

Figura 15 – Exemplo de *dashboard* criada com o Spago BI.

Fonte: <https://www.youtube.com/watch?v=of9nPTrY1Pc>. Acesso em 05 de março de 2018

4.2.2.2.2 Ferramentas de ETL e processamento dos dados

Para o conjunto de ferramentas específicas é necessária a utilização de uma ferramenta que permita acessar o dado, limpá-lo, além de processá-lo através de diferentes técnicas. Para este fim foram analisadas duas ferramentas: o Hadoop e o Spark. O Hadoop é uma das plataformas de destaque na escalabilidade horizontal, isto é, na distribuição do trabalho em múltiplas máquinas de forma a melhorar a capacidade de processamento. O Spark vêm sido desenvolvido como parte da próxima geração deste tipo de plataforma, buscando superar as limitações das outras plataformas (SINGH; REDDY, 2015).

Hadoop

O Hadoop é um *framework open-source* da Apache voltado para o armazenamento de dados e processamento *batch* utilizando *clusters*. O HDFS (*Hadoop distributed file system*) é o sistema de arquivos utilizado pelo Hadoop para armazenamento dos dados. Neste sistema os dados são distribuídos em múltiplos nós contidos nos *clusters*. O HDFS

utiliza uma arquitetura mestre-escravo, na qual os *datanodes* são os nós escravos, e o *namenode* é o nó mestre. Os *datanodes* armazenam os dados em forma de blocos e, sob demanda, reportam ao *namenode* informações sobre estes dados. O *namenode* gerencia os arquivos, mantendo a referência de suas localizações e direcionando o tráfego para os *datanodes*. Este sistema é tolerante à falhas, uma vez que várias cópias dos blocos de dadas são armazenadas para caso ocorram falhas no disco.

O MapReduce é o esquema básico de processamento de dados utilizado no Hadoop, nele cada *job* possui duas fases executadas paralelamente no *cluster*: mapeamento e redução. No mapeamento, os dados são obtidos e organizados em pares chave-valor. Na redução, os dados são processados em paralelo e agregados, de forma a gerar uma saída final.

Um dos problemas do Hadoop é o frequente acesso ao disco. Em processos iterativos, depois de cada iteração, os dados precisam ser escritos no disco para serem passados para a próxima iteração, tornando o processo ineficiente nestes casos (SINGH; REDDY, 2015), (LANDSET et al., 2015).

Spark

O Spark também é uma ferramenta de processamento distribuído, do mesmo ecossistema do Hadoop. Ele suporta as linguagens Java, Scala, Python e R. A sua principal característica em relação ao Hadoop é a utilização da memória ao invés do disco, eliminando a limitação para tarefas iterativas que o Hadoop apresenta. Para isto, o Spark utiliza a estratégia de *micro-batches*. É possível definir o tempo de processamento dos *micro-batches* em código, permitindo que sejam definidos tempos pequenos o suficiente para serem considerados tempo real. Além dessa vantagem, o Spark ainda permite o processamento *streaming* e possui bibliotecas nativas de aprendizado de máquina (SINGH; REDDY, 2015), (GUEDES, 2017).

Considerando a performance superior ao Hadoop e a disponibilidade de bibliotecas nativas de aprendizado de máquina, o Spark foi selecionado. Sua utilização se dará tanto no processo de ETL, isto é, na leitura, carregamento e transformação dos dados, quanto no processamento dos dados, através principalmente de *queries* e aprendizado de máquina.

4.2.2.2.3 Ferramentas de visualização dos dados

Para compor o conjunto de ferramentas específicas, também é necessária a utilização de uma ferramenta que permita a visualização dos dados processados. Considerando que as ferramentas de BI escaláveis avaliadas são do ecossistema Hadoop/Spark, optou-se por avaliar as ferramentas de visualização do mesmo ecossistema.

Dentre opções como o Solr, Lens e Zeppelin, o Zeppelin foi a ferramenta que mostrou melhor aderência ao Spark. O Zeppelin possui vários "intérpretes"embutidos, que podem interpretar e invocar todas as funções da API do Spark, inclusive as bibliotecas de aprendizado de máquina e o SparkSQL, que mapeia os dados de entrada em *queries* SQL, permitindo por exemplo a implementação de um ROLAP (*Relational On Line Analytical Processing*). Além disso, o Zeppelin também suporta Python, Angular, Shell, dentre outras tecnologias, e há bibliotecas disponíveis para vários propósitos (CHITTURI, 2016).

Uma das principais vantagens é que a ferramenta é altamente customizável, podendo ser definidos vários blocos de código para serem executados e exibidos através de diversos tipos de gráficos, também customizáveis, conforme ilustra a Figura 16, mostrando as possíveis configurações para uma *query* de duas colunas. Além disso, o Zeppelin também possui uma simples configuração, decorrente da integração embutida com o Spark.

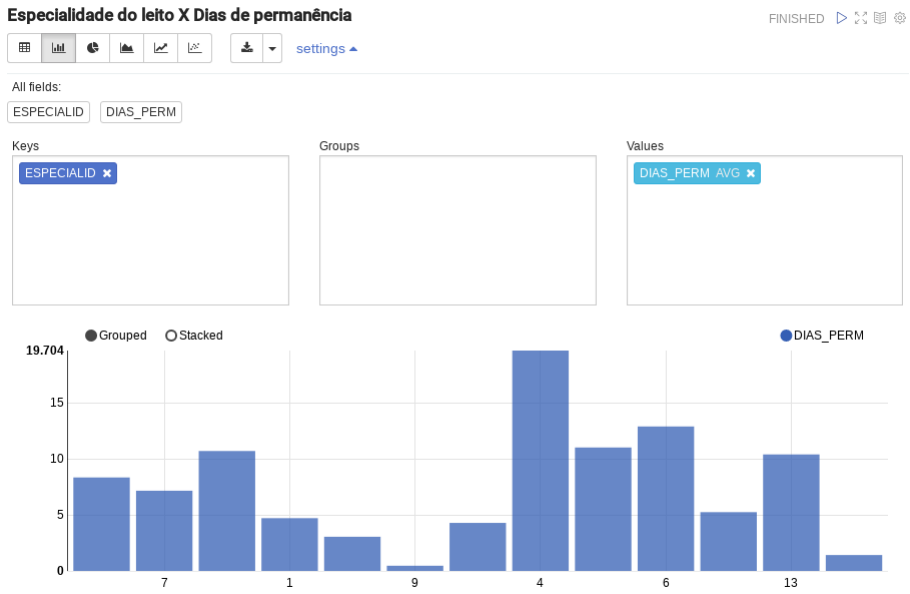

Figura 16 – Possíveis configurações gráficas para uma *query* no Zeppelin.

4.2.2.2.4 Solução Final

Apesar do Pentaho e do Spago BI serem ferramentas muito completas, sua utilização não é tão simples, pois a instalação/configuração destas ferramentas envolvem diversos módulos. Quando comparada à instalação/configuração do Zeppelin com o Spark, esta segunda opção é mais prática neste contexto, levando à escolha destas ferramentas. Além disso, outro ponto contribuinte foi a alta customização do Zeppelin com o Spark, uma vez que a dificuldade de customização é um problema crítico das ferramentas utilizadas atualmente no contexto estudado. Por fim, o Zeppelin mostrou uma maior usabilidade em relação às outras ferramentas, permitindo que, uma vez configurado, pessoas sem muito

domínio da ferramenta consigam utilizá-lo para manipulação dos gráficos, execução de *jobs*, dentre outros. A Figura 17 ilustra a estrutura da solução final.

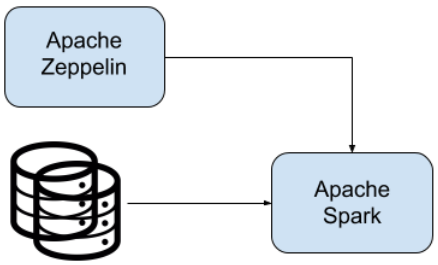

Figura 17 – Ferramentas da solução final.

A ferramenta de visualização Zeppelin provê um ambiente em que o código Spark pode ser desenvolvido e executado. Através dela, e utilizando código Spark, é possível ler dados tanto da nuvem, quanto locais. Estes dados são armazenados nos chamados RDDs (*Resilient Distributed Dataset*), que são coleções imutáveis e distribuídas dos dados. Os RDDs podem ainda ser acessados por *DataFrames*, eles fornecem uma API de linguagem específica do domínio para manipular os dados distribuídos, organizando-os em colunas. Além disso, é possível o acesso também por *queries* SQL, através do SparkSQL.

4.3 Resultados Parciais

Neste capítulo serão apresentados os principais resultados alcançados até o momento, conforme o fluxo de atividades planejado para este trabalho. Os resultados estão descritos na seção de Identificação dos dados (seção 4.3.1) e Validação das ferramentas (seção 4.3.2), além da seção Ações Futuras (4.3.3), onde são descritos os próximos passos após este trabalho.

4.3.1 Identificação dos dados

A Secretaria da Saúde de São Paulo (SES-SP) disponibilizou alguns de seus dados para utilização no projeto. Foi realizada uma reunião com a Secretaria da Saúde do Distrito Federal para verificar se a mesma também possuía as informações disponibilizadas pela SES-SP.

A tabela no Apêndice A descreve os tipos dos dados disponibilizados, sendo principalmente informações do paciente, informações do estabelecimento de atendimento do paciente, e informações sobre a relação do paciente e do estabelecimento, isto é, diárias, especialidade do leito, diagnóstico, dentre outros.

As informações de diagnóstico são mapeadas com o CID 10 (Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde), que fornece

códigos relativos à classificação de doenças, sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças.

4.3.2 Validação das ferramentas

A fim de validar as ferramentas selecionadas, foi criada uma aplicação simples utilizando os dados da SES-SP e abordagens básicas de análise. Nesta aplicação, foram usadas tanto *queries* e gráficos, quanto o aprendizado de máquina, na tentativa de buscar informações quanto a relação entre a especialidade do leito e os dias de permanência do paciente no leito. A Figura 18 ilustra os resultados obtidos através destas duas abordagens.

Figura 18 – *Dashboard* de BI utilizando o Zeppelin para análise de leitos hospitalares.

4.3.2.1 Configuração e Instalação

É possível tanto realizar o *download* do Zeppelin, quanto executá-lo em um *container*. Para esta aplicação, foi utilizado um *container* Docker com a imagem oficial disponibilizada pela Apache.

É possível fazer a configuração do intérprete Spark, ou até mesmo de outros intérpretes, utilizando a própria interface do Zeppelin (Figura 19), ou através de variáveis de ambiente.

4.3.2.2 Processamento dos dados

O Spark permite a leitura de dados de diferentes fontes e em diferentes formatos. Os dados utilizados nesta aplicação são do formato Shapefile. Foi utilizada a biblioteca

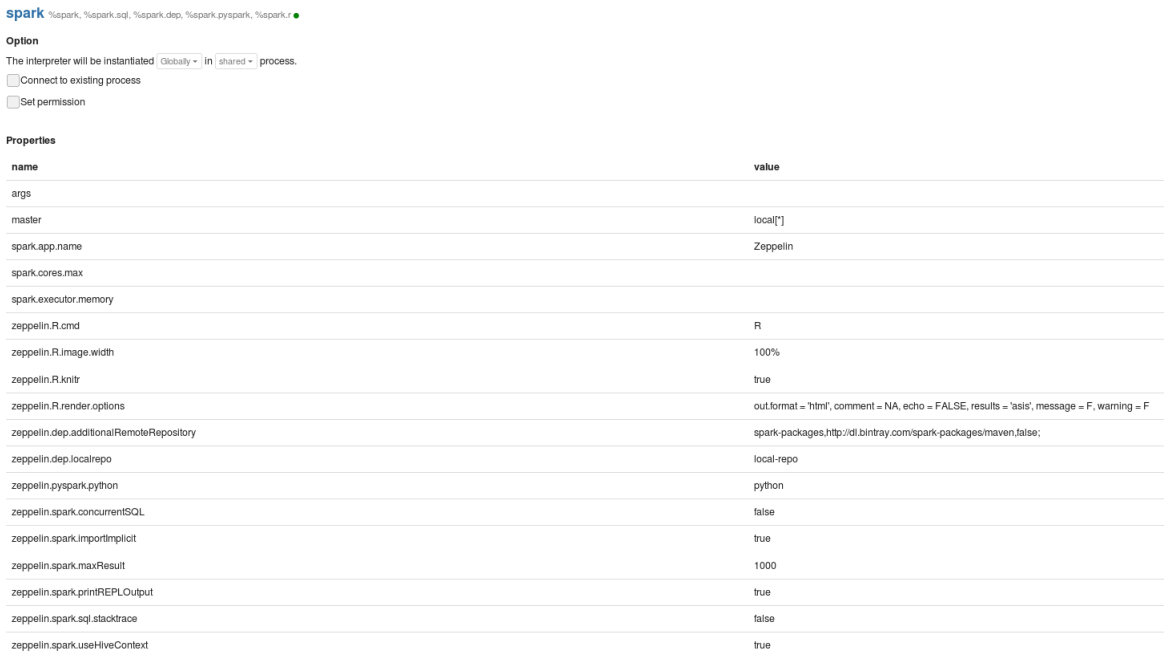

Figura 19 – Página de configuração do intérprete Spark no Zeppelin.

Magellan para conversão dos dados deste formato, para o DataFrame do Spark. Posteriormente foi criada uma tabela temporária a partir deste DataFrame, possibilitando a execução de *queries* SQL. Este processo está ilustrado na Figura 20.

```
%spark
import magellan.{Point, Polygon, PolyLine}

val data = sqlContext.read.format("magellan").load("data")
data.show(3)
data.createOrReplaceTempView("procedure")
```

```
import magellan.{Point, Polygon, PolyLine}
data: org.apache.spark.sql.DataFrame = [point: point, polyline: polyline ... 3 more fields]
+--------------------+--------+-------+--------------------+-----+
|               point|polyline|polygon|            metadata|valid|
+--------------------+--------+-------+--------------------+-----+
|Point(-46.784632,...|    null|   null|Map(DIAG_SE3 ->  ...| true|
|Point(-46.708909,...|    null|   null|Map(DIAG_SE3 ->  ...| true|
|Point(-46.459507,...|    null|   null|Map(DIAG_SE3 ->  ...| true|
+--------------------+--------+-------+--------------------+-----+
only showing top 3 rows
```

Figura 20 – Processamento do arquivo Shapefile no Spark.

Com os dados processados, foi possível utilizá-los para as análises. As Figuras 21, 23, 22 ilustram algumas das possíveis visualizações para uma *query* de seleção de duas colunas: especialidade do leito e dias de permanência do paciente no leito. O gráfico de barras da Figura 21 representa a média de dias de permanência por leito, através desta visualização é possível ter uma visão de quais leitos são ocupados por mais ou menos dias, entretanto, ainda é possível reconfigurar o gráfico para visualização da soma, valor máximo, mínimo e a quantidade de registros. Esta representação também pode ser feita

utilizando o gráfico de pizza (Figura 22).

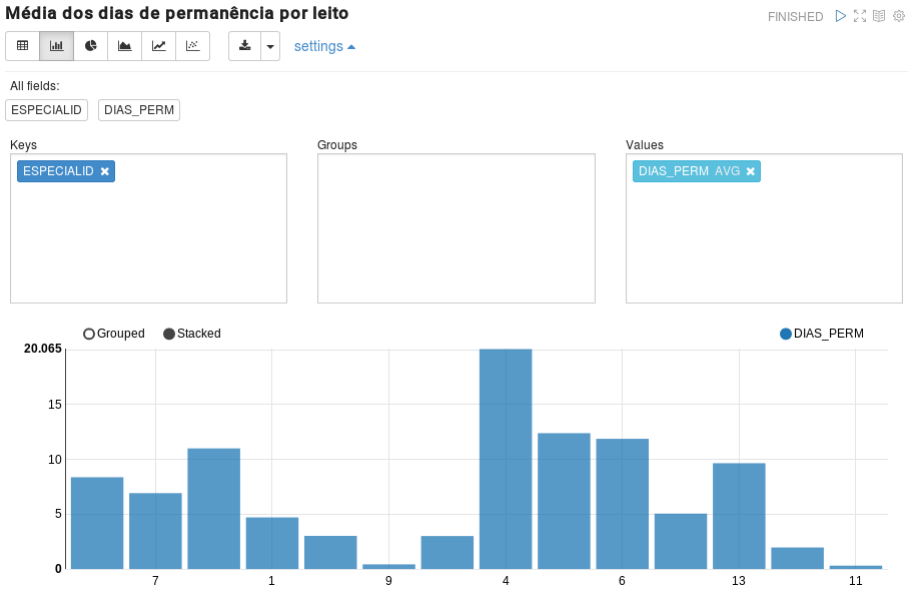

Figura 21 – Visualização do gráfico de barras representando a média de dias de permanência por leito.

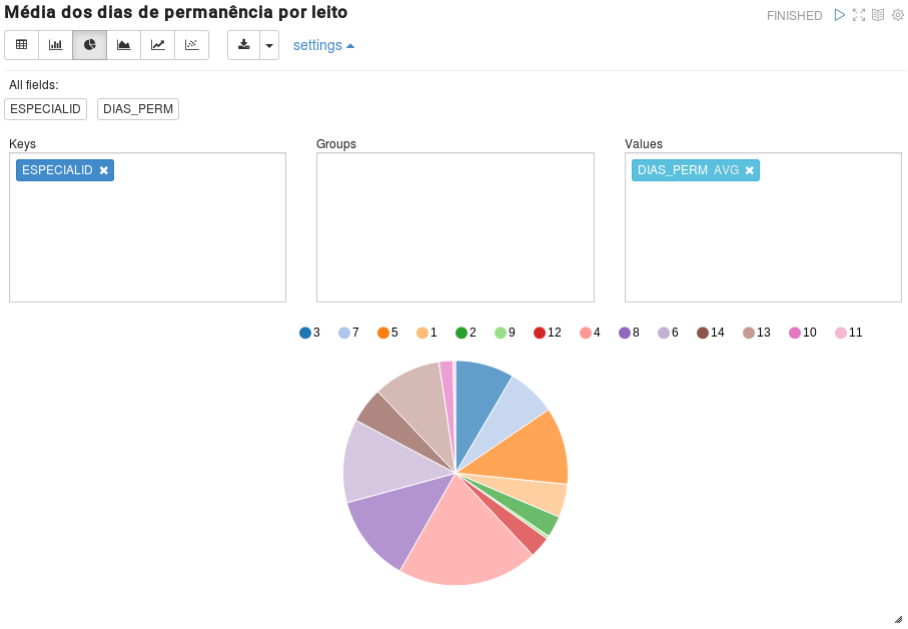

Figura 22 – Visualização do gráfico de pizza representando a média de dias de permanência por leito.

É possível modificar o gráfico para representar a dispersão dos dados, possibilitando que uma pessoa analise-o e identifique leitos que possuem anomalias, isto é, registros de dias de permanência muito altos ou muito baixos em relação ao que é comum para aquele leito (Figura 23).

Figura 23 – Visualização do gráfico de dispersão relacionando a especialidade do leito e os dias de permanência do paciente no leito.

Através do aprendizado de máquina é possível a utilização de algoritmos que aprendam a detectar estas anomalias, desta forma, quando novos registros entrarem no sistema, o algoritmo consegue identificar se eles são ou não anomalias, tornando a análise mais prática. Esta tentativa foi realizada nesta aplicação através da integração do Python com o Zeppelin, utilizando a implementação do algoritmo K-Means da biblioteca scikit-learn. O resultado é ilustrado na Figura 24, onde foram detectadas sete anomalias cujas informações foram descritas na Figura 25.

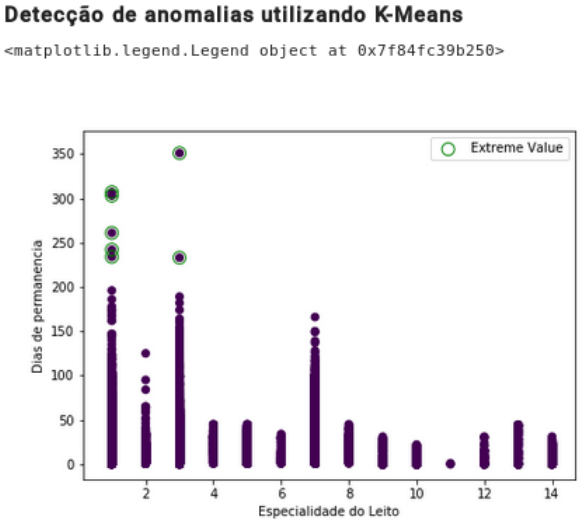

Figura 24 – Identificação de anomalias nos dias de permanência dos leitos utilizando o algoritmo K-Means.

Informações sobre as anomalias encontradas

Raça	Especialidade	Idade	Dias de permanência no leito	Diagnóstico Principal
Branca	Clínica Médica	81 anos	351 dias	Algumas doenças infecciosas e parasitárias
Sem Informação	Cirurgia	49 anos	307 dias	Lesões, envenenamento e algumas outras conseqüências de causas externas
Sem informação	Cirurgia	8 anos	303 dias	Doenças do aparelho digestivo
Branca	Cirurgia	49 anos	261 dias	Doenças do aparelho circulatório
Branca	Cirurgia	0 anos	242 dias	Doenças do sistema nervoso
Sem Informação	Cirurgia	65 anos	234 dias	Doenças do aparelho circulatório
Branca	Clinica Médica	37 anos	233 dias	Algumas doenças infecciosas e parasitárias

Figura 25 – Informações sobre as anomalias identificadas.

A Figura 26 ilustra a tentativa de encontrar padrões nos dados de dias de permanência e idade, no leito de cirurgia. O gráfico à esquerda representa a dispersão dos dados, e o gráfico à direita, representa a tentativa de classificação destes dados em grupos semelhantes através do algoritmo K-Means.

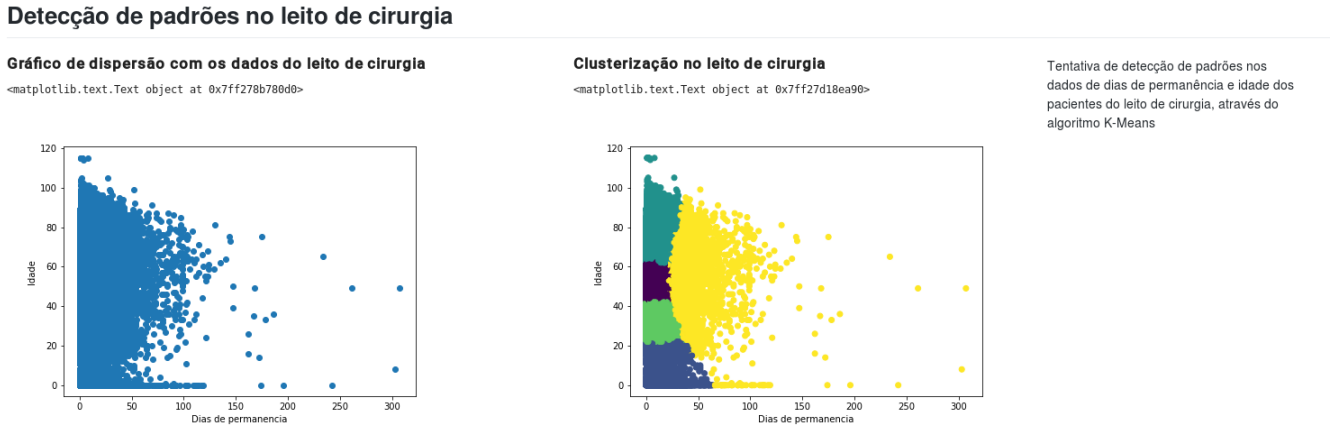

Figura 26 – Detecção de padrões no leito de cirurgia através do algoritmo K-Means.

Com os resultados obtidos não parece existir uma relação entre as características analisadas. Este resultado evidencia a relevância da pesquisa, uma vez que nem sempre o aprendizado de máquina será tão útil quanto ou melhor que as estratégias clássicas. Além disso, mostra a importância da fase de definição dos objetivos e do envolvimento de pessoas com conhecimento do domínio, uma vez que características que parecem relacionadas, podem não ter relação no contexto analisado. Outro ponto importante é que como foram adotadas estratégias básicas para validação das ferramentas, é possível que explorando outros tipos de algoritmos, como os supervisionados, ou com um processo mais complexo de engenharia de *feature*, obtenham-se melhores resultados.

5 Resultados Finais

Este capítulo tem o objetivo de descrever os resultados obtidos na aplicação da metodologia descrita na seção 4.1.

5.1 Objetivo 1: Previsão de demandas de internações

O primeiro objetivo definido foi a previsão da quantidade de demandas de internações. Para isto foram selecionadas inicialmente duas *features*: as datas das internações e a quantidade de internações para cada data. Estes dados são rotulados, caracterizando o modelo como de caráter supervisionado, e como os dados possuem valores contínuos, o problema caracteriza-se como de regressão.

5.1.1 Processando os dados

Através da visualização do gráfico da Figura 27 que relaciona as datas e as quantidades de internações, foi possível identificar uma inconsistência nos dados do ano de 2014 e uma queda no final do ano de 2015. Eliminou-se as inconsistências e os dados de 24 à 31 de dezembro de 2015, uma vez que não há informações sobre os outros anos para identificar se a queda é ou não um padrão, desta forma, obteve-se o gráfico da Figura 28.

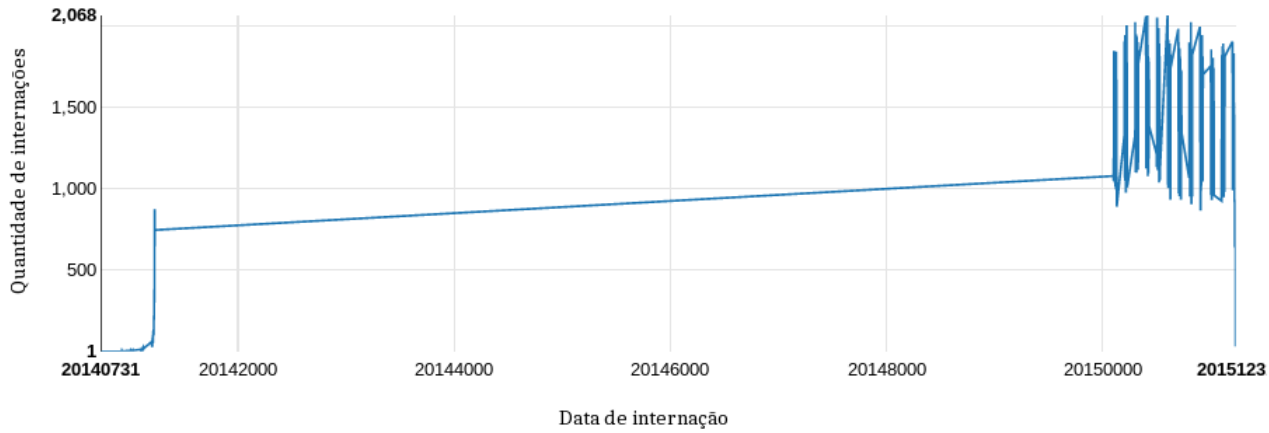

Figura 27 – Gráfico que relaciona a quantidade de internações e as datas das internações para os anos de 2014 e 2015.

5.1.2 Estratégias de análise

Para o objetivo descrito na seção 5.1 foram comparados dois algoritmos: a *Random Forest* e a classe de rede neural LSTM (*Long Short-Term Memory*).

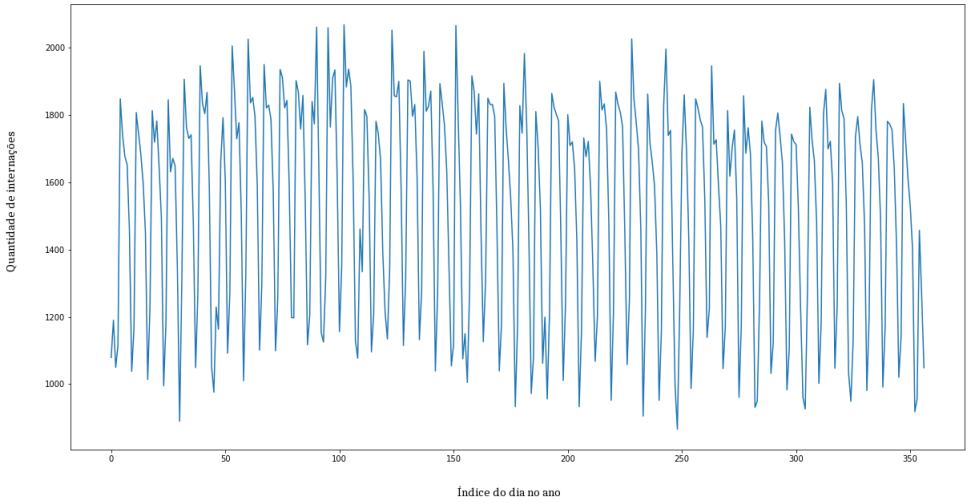

Figura 28 – Gráfico que relaciona a quantidade de internações e as datas das internações para os anos de 2015.

5.1.3 Long Short-Term Memory

A LSTM é um tipo de RNN (*Recurrent Neural Network*), nas RNNs a relação temporal das séries pode ser representada usando *loops* de *feedback* para os nós internos da camada oculta, dessa forma é possível propagar informação fazendo o mapeamento entre as sequências de entrada e saída das redes. O problema da RNN tradicional é o *vanishing gradient problem*, isto é, a influência de uma entrada na camada oculta, cai exponencialmente à medida que esta circula pelas conexões recorrentes da rede. Para mitigar este problema a LSTM substiui ou adiciona blocos de memória na camada oculta, cada bloco contém células de memória auto-conectadas e três unidades multiplicativas - as portas de entrada, saída e de "esquecimento", as portas multiplicativas permitem que as células de memória armazenem e acessem informações por longos períodos de tempo (CHENG et al., 2006), (KAWAKAMI, 2008).

5.1.3.1 Construção do modelo

Para a LSTM, os dados foram formatados no formato X=t e Y=t+1. Além disso, foi feita a diferença do vetor a fim de remover sistematicidades, como tendências e sazonalidades, que podem dificultar o desempenho do modelo (DORFFNER, 1996). Após este processo obteve-se o vetor da Figura 29. Por fim, os dados foram normalizados, mudando a escala para o *range* de -1 a 1, e foram divididos em treino e teste, sendo setenta por cento dos dados dedicados ao treino e trinta por cento para teste.

O modelo contém uma camada visível, uma camada oculta com quatro blocos de memória e uma camada de saída. No resultado da Figura 30 a rede foi treinada com três mil épocas e com um *batch size* de um, obtendo um erro absoluto médio percentual de 0.095.

	X	Y
0	0.0	111
1	111.0	-140
2	-140.0	66
3	66.0	732
4	732.0	-107
5	-107.0	-64
6	-64.0	-25
7	-25.0	-196
8	-196.0	-418
9	-418.0	126
10	126.0	643
11	643.0	-63
12	-63.0	-61
13	-61.0	-83
14	-83.0	-148
15	-148.0	-438

Figura 29 – Primeiros 15 itens do vetor dos dados de entrada após reformatação e diferenciação.

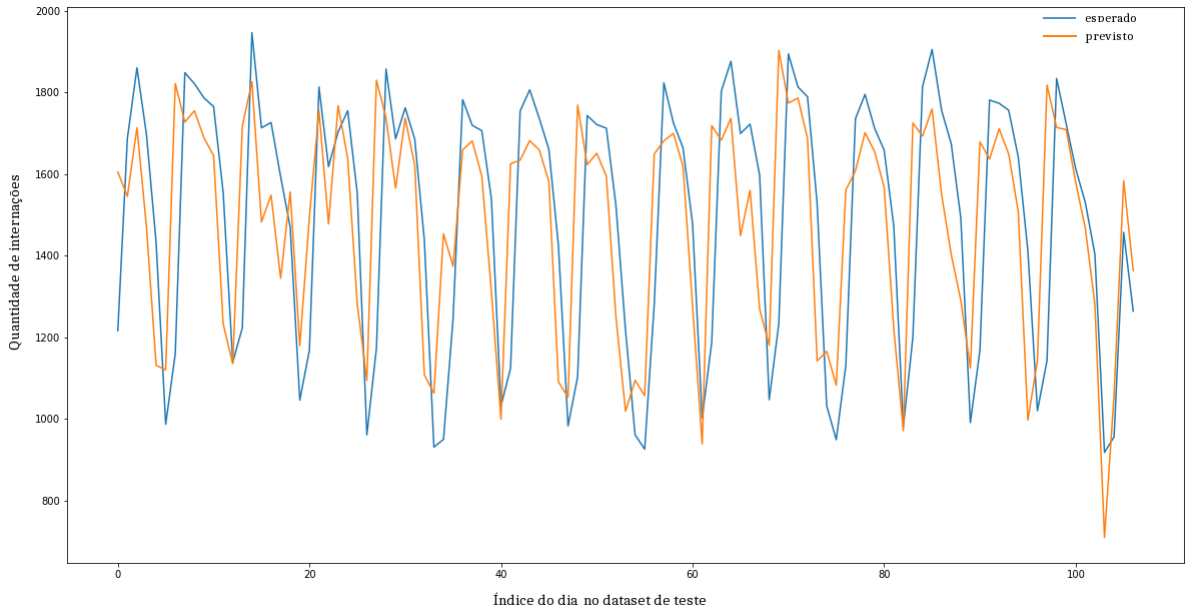

Figura 30 – Valores esperados e previstos através da LSTM para os dados de teste

5.1.4 Random Forest

O *Random Forest* é um algoritmo que pode ser usado tanto para classificação como para regressão, para isto, ele utiliza um conjunto de árvores de decisão criadas a partir da divisão aleatória dos dados de treino em conjuntos menores, então, a predição é feita por todas as árvores e seus resultados são combinados no final por média ou pela maioria (BREIMAN, 2001). Segundo Tyralis e Papacharalampous (2017), este método se mostrou eficiente para a predição de séries temporais, além de ser de uso simples, ser eficiente para

dados de pequeno tamanho e possuir poucos parâmetros para otimização.

5.1.4.1 Construção do modelo

Nos dados foi possível observar um padrão de em média quatro quedas por mês nas quantidades de internações, levantando a hipótese das quedas serem aos finais de semana. Dessa forma, para a *Random Forest* foi adicionada uma nova *feature* aos dados, chamada "*dayofweek*", um índice de 0 à 6, onde 0 representa as segunda-feiras e 6 representa os domingos.

A Figura 31 ilustra o resultado da *Random Forest* com árvores com uma profundidade máxima de cinco e as mesmas configurações de tamanho de treino e teste usadas na LSTM, obtendo um erro absoluto médio percentual de 0.079.

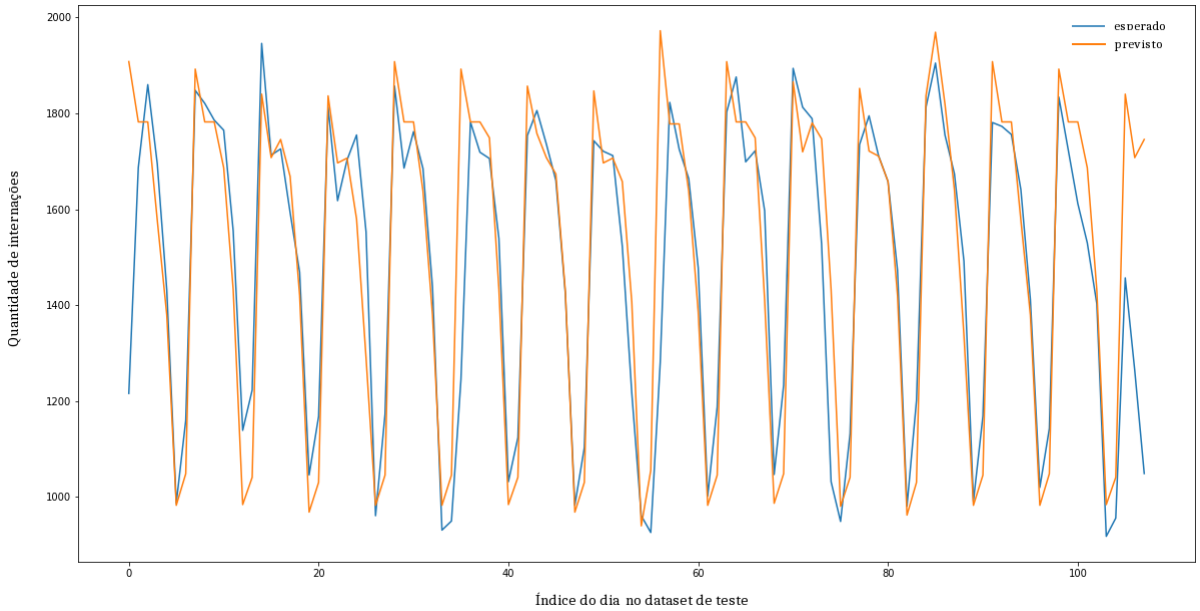

Figura 31 – Valores esperados e previstos através da *Random Forest* para os dados de teste

Para verificar se a *feature* adicionada foi realmente relevante, utilizou-se do atributo *feature importances* da biblioteca *scikit learn*. A Figura 32 ilustra que a *feature* adicionada foi a mais relevante para o algoritmo.

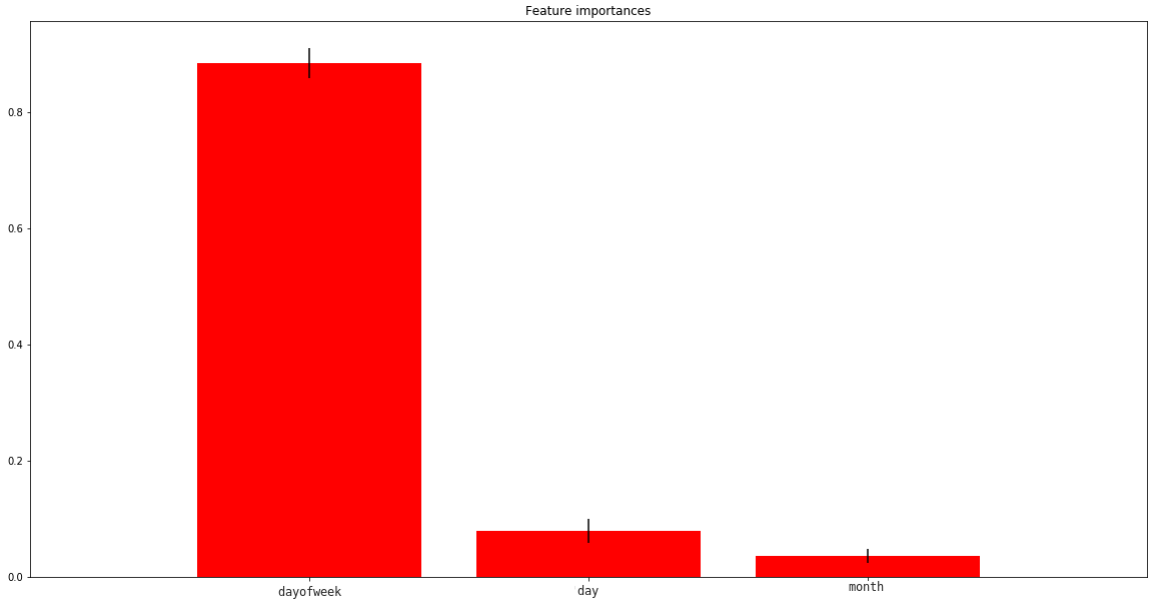

Figura 32 – Relevância das *features* para o resultado da *Random Forest*

6 Conclusão

O aprendizado de máquina se mostrou uma estratégia adequada na previsão de demandas de internações no contexto analisado. Ambos os algoritmos mostraram desempenho satisfatório, sendo que a *Random Forest* apresentou um erro menor que a LSTM, isto é, a *Random Forest* apresentou um erro absoluto médio percentual de 0.079 e a LSTM, de 0.095. Apesar da *Random Forest* ser um algoritmo mais básico, ele se mostra eficiente para conjuntos de dados menores e não exige uma otimização complexa dos parâmetros (TYRALIS; PAPACHARALAMPOUS, 2017), o que pode justificar o resultado, uma vez que foram utilizados dados de apenas um ano e que as configurações dos parâmetros foram testadas manualmente, técnicas como o *Grid Search* podem melhorar o desempenho da LSTM.

Outro ponto importante é que a biblioteca utilizada para a LSTM, o Keras, utiliza da randomicidade para a inicialização dos pesos da rede, dessa forma, a mesma configuração pode gerar diferentes resultados. Oerro mínimo obtido neste experimento para a LSTM foi de 0.077, entretanto, a média do erro de múltiplas execuções da mesma configuração ainda é maior que o erro da *Random Forest*.

Quanto às ferramentas tanto o Spark como o Zeppelin se mostraram eficientes no processamento dos dados e na visualização dos mesmos. A plataforma Zepl permite a importação de *Zeppelin Notebooks* de forma que eles podem ser publicados e configurados para serem executados automaticamente em determinados intervalos de tempo, sendo uma opção rápida e prática para o *deploy*.

Referências

AALST, W. M. Van der. Data scientist: The engineer of the future. In: *Enterprise Interoperability VI*. [S.l.]: Springer, 2014. p. 13–26. Citado na página 23.

ALNOUKARI, M.; RAZOUK, R.; HANANO, A. Bsc-si: A framework for integrating strategic intelligence in corporate strategic management. *International Journal of Social and Organizational Dynamics in IT (IJSODIT)*, IGI Global, v. 5, n. 2, p. 1–14, 2016. Citado na página 16.

ALSAFFAR, M. et al. The state of open source electronic health record projects: A software anthropology study. *JMIR medical informatics*, JMIR Publications Inc., v. 5, n. 1, 2017. Citado na página 21.

AYANKOYA, K.; CALITZ, A.; GREYLING, J. Intrinsic relations between data science, big data, business analytics and datafication. In: ACM. *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology*. [S.l.], 2014. p. 192. Citado 2 vezes nas páginas 9 e 25.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. *Artificial intelligence*, Elsevier, v. 97, n. 1, p. 245–271, 1997. Citado na página 30.

BOSE, I.; MAHAPATRA, R. K. Business data mining—a machine learning perspective. *Information & management*, Elsevier, v. 39, n. 3, p. 211–225, 2001. Citado 2 vezes nas páginas 16 e 29.

BRANDÃO, A. et al. A benchmarking analysis of open-source business intelligence tools in healthcare environments. *Information*, Multidisciplinary Digital Publishing Institute, v. 7, n. 4, p. 57, 2016. Citado 3 vezes nas páginas 37, 39 e 40.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 51.

CASOLA, V. et al. Healthcare-related data in the cloud: Challenges and opportunities. *IEEE Cloud Computing*, IEEE, v. 3, n. 6, p. 10–14, 2016. Citado 3 vezes nas páginas 9, 19 e 20.

CHAUDHURI, S.; DAYAL, U.; GANTI, V. Database technology for decision support systems. *Computer*, IEEE, v. 34, n. 12, p. 48–55, 2001. Citado 2 vezes nas páginas 9 e 28.

CHAUDHURI, S.; DAYAL, U.; NARASAYYA, V. An overview of business intelligence technology. *Communications of the ACM*, ACM, v. 54, n. 8, p. 88–98, 2011. Citado 5 vezes nas páginas 9, 26, 27, 28 e 29.

CHENG, H. et al. Multistep-ahead time series prediction. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2006. p. 765–774. Citado na página 50.

CHITTURI, P. *Apache Spark for Data Science Cookbook*. Packt Publishing, 2016. ISBN 9781785288807. Disponível em: <https://books.google.com.br/books?id=hdDcDgAAQBAJ>. Citado na página 42.

COSTA, C. G. A. d. et al. Construção de um ambiente de bi (business intelligence) na secretaria municipal de saúde da cidade de são paulo. 2008. Citado na página 22.

DELEN, D.; DEMIRKAN, H. *Data, information and analytics as services*. [S.l.]: Elsevier, 2013. Citado 3 vezes nas páginas 9, 24 e 26.

DHAR, V. Data science and prediction. *Communications of the ACM*, ACM, v. 56, n. 12, p. 64–73, 2013. Citado na página 23.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, v. 55, n. 10, p. 78–87, 2012. Citado na página 30.

DORFFNER, G. Neural networks for time series processing. In: CITESEER. *Neural network world*. [S.l.], 1996. Citado na página 50.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 1996. Citado na página 16.

FERRAZ, R. N. Uma solução de business intelligence baseada em data warehouse para a secretaria de saúde do estado de pernambuco. Universidade Federal de Pernambuco, 2009. Citado na página 21.

GONÇALVES, D.; SANTOS, M. Y.; CRUZ, J. M. Implementação de um sistema de business intelligence para a análise da qualidade de vida pré e pós-operatória. *Sistemas e Tecnologias de Informação na Saúde*, Edições Universidade Fernando Pessoa, p. 93–110, 2010. Citado na página 22.

GUEDES, D. J. M. G. Processamento de dados em uma plataforma de cidades inteligentes. 2017. Citado na página 41.

HAN, J. Olap mining: An integration of olap with data mining. In: *Proceedings of the 7th IFIP*. [S.l.: s.n.], 1997. v. 2, p. 1–9. Citado na página 28.

HAUX, R. Health information systems–past, present, future. *International journal of medical informatics*, Elsevier, v. 75, n. 3, p. 268–281, 2006. Citado 2 vezes nas páginas 16 e 19.

KAWAKAMI, K. *Supervised sequence labelling with recurrent neural networks*. Tese (Doutorado) — Ph. D. thesis, Technical University of Munich, 2008. Citado na página 50.

KIMBALL, R.; CASERTA, J. *The Data Warehouse? ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. [S.l.]: John Wiley & Sons, 2011. Citado na página 26.

KOH, H. C.; TAN, G. et al. Data mining applications in healthcare. *Journal of healthcare information management*, v. 19, n. 2, p. 65, 2011. Citado na página 22.

LANDSET, S. et al. A survey of open source tools for machine learning with big data in the hadoop ecosystem. *Journal of Big Data*, Springer, v. 2, n. 1, p. 24, 2015. Citado na página 41.

LANS, R. V. D. *Data Virtualization for business intelligence systems: revolutionizing data integration for data warehouses*. [S.l.]: Elsevier, 2012. Citado na página 27.

LAPA, J.; BERNARDINO, J.; FIGUEIREDO, A. A comparative analysis of open source business intelligence platforms. In: ACM. *Proceedings of the International Conference on Information Systems and Design of Communication*. [S.l.], 2014. p. 86–92. Citado 2 vezes nas páginas 37 e 39.

LAROSE, D. T. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2014. Citado na página 31.

LARSON, D.; CHANG, V. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, Elsevier, v. 36, n. 5, p. 700–710, 2016. Citado na página 36.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Citado na página 30.

MAGLOGIANNIS, I. G. *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*. [S.l.]: Ios Press, 2007. v. 160. Citado na página 30.

MARSLAND, S. *Machine learning: an algorithmic perspective*. [S.l.]: CRC press, 2015. Citado 3 vezes nas páginas 29, 30 e 31.

METTLER, T.; VIMARLUND, V. Understanding business intelligence in the context of healthcare. *Health informatics journal*, Sage Publications Sage UK: London, England, v. 15, n. 3, p. 254–264, 2009. Citado na página 15.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, v. 1, n. 1, 2003. Citado 4 vezes nas páginas 29, 30, 31 e 32.

NEGASH, S. Business intelligence. *The communications of the Association for Information Systems*, v. 13, n. 1, p. 54, 2004. Citado na página 15.

OESTREICH, T. W. Magic quadrant for business intelligence and analytics platforms. *Analyst (s)*, v. 501, p. G00275847, 2016. Citado 2 vezes nas páginas 37 e 39.

PAIXÃO, A. d. O.; SILVA, V. A. d.; TANAKA, A. De business intelligence a data science: um estudo comparativo entre áreas de conhecimento relacionadas. *VIII Congresso Integrado de Tecnologia da Informação, Campos dos Goytacazes, RJ*, 2015. Citado na página 23.

PORTO, F.; ZIVIANI, A. Ciência de dados. *III Seminário de Grandes Desafios da Computação no Brasil, Rio de Janeiro, RJ*, 2014. Citado na página 23.

PRODANOV, C. C.; FREITAS, E. C. de. *Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico-2ª Edição*. [S.l.]: Editora Feevale, 2013. Citado 3 vezes nas páginas 9, 33 e 34.

PROVOST, F.; FAWCETT, T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. [S.l.]: "O'Reilly Media, Inc.", 2013. Citado na página 23.

RAGHUPATHI, W.; RAGHUPATHI, V. Big data analytics in healthcare: promise and potential. *Health information science and systems*, BioMed Central, v. 2, n. 1, p. 3, 2014. Citado 5 vezes nas páginas 15, 25, 26, 27 e 38.

REYNOLDS, C. J.; WYATT, J. C. Open source, open standards, and health care information systems. *Journal of medical Internet research*, JMIR Publications Inc., v. 13, n. 1, 2011. Citado na página 21.

ROUHANI, S.; ASGARI, S.; MIRHOSSEINI, S. V. Review study: business intelligence concepts and approaches. *American Journal of Scientific Research*, v. 50, n. 1, p. 62–75, 2012. Citado na página 25.

SAFADI, H. et al. Open-source health information technology: A case study of electronic medical records. *Health Policy and Technology*, Elsevier, v. 4, n. 1, p. 14–28, 2015. Citado 2 vezes nas páginas 20 e 21.

SANTOS, R. F. dos. Estruturação de um ambiente de business intelligence (bi) para gestão da informação em saúde: a experiência da secretaria municipal de saúde de belo horizonte. *Journal of Health Informatics*, v. 3, n. 4, 2011. Citado na página 21.

SCHWABER, K.; SUTHERLAND, J. The scrum guide-the definitive guide to scrum: The rules of the game, july 2011. *Available on-line at: http://www. scrum. org/storage/scrumguides/Scrum% 20Guide*, 2016. Citado na página 36.

SINGH, D.; REDDY, C. K. A survey on platforms for big data analytics. *Journal of Big Data*, Springer, v. 2, n. 1, p. 8, 2015. Citado 2 vezes nas páginas 40 e 41.

SOLANAS, A. et al. Smart health: a context-aware health paradigm within smart cities. *IEEE Communications Magazine*, IEEE, v. 52, n. 8, p. 74–81, 2014. Citado 2 vezes nas páginas 15 e 19.

SONG, I.-Y.; ZHU, Y. Big data and data science: what should we teach? *Expert Systems*, Wiley Online Library, v. 33, n. 4, p. 364–373, 2016. Citado na página 23.

STANTON, J. M. Introduction to data science. 2013. Citado na página 23.

TAVARES, L. G.; LOPES, H. S.; LIMA, C. R. E. Estudo comparativo de métodos de aprendizado de máquina na detecção de regiões promotoras de genes de escherichia coli. *Anais do I Simpósio Brasileiro de Inteligência Computacional*, p. 8–11, 2007. Citado na página 31.

TYRALIS, H.; PAPACHARALAMPOUS, G. Variable selection in time series forecasting using random forests. *Algorithms*, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 114, 2017. Citado 2 vezes nas páginas 51 e 55.

WASHBURN, D. et al. Helping cios understand "smart city" initiatives. *Growth*, v. 17, n. 2, p. 1–17, 2009. Citado na página 15.

WATSON, H. J.; WIXOM, B. H. The current state of business intelligence. *Computer*, IEEE, v. 40, n. 9, 2007. Citado 3 vezes nas páginas 25, 27 e 38.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado 3 vezes nas páginas 29, 30 e 32.

World Bank Group. *The United Nations Population Division's World Urbanization Prospects*. 2016. <https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS>. Accessed: 2018-01-05. Citado na página 15.

YELLOWLEES, P. M. et al. Standards-based, open-source electronic health record systems: a desirable future for the us health industry. *Telemedicine and e-Health*, Mary Ann Liebert, Inc. 140 Huguenot Street 3rd Floor New Rochelle, NY 10801 USA, v. 14, n. 3, p. 284–288, 2008. Citado 2 vezes nas páginas 20 e 21.

ZUMEL, N.; MOUNT, J.; PORZAK, J. *Practical data science with R*. [S.l.]: Manning, 2014. Citado 3 vezes nas páginas 9, 24 e 35.

Apêndices

APÊNDICE A – Identificação dos dados

Tabela 1 – Dados da Secretaria de Saúde de São Paulo.

Categoria	**Descrição**
Paciente	Sexo
	Idade
	Código da Raça/Cor
	Nível de Instrução
	Código do setor censitário da residência
	Coordenada geográfica do centroide do setor censitário (latitude)
	Coordenada geográfica do centroide do setor censitário (longitude)
Estabelecimento	Coordenada geográfica do centroide do estabelecimento de saúde (latitude)
	Coordenada geográfica do centroide do estabelecimento de saúde (longitude)
	Cadastro Nacional de Estabelecimento de Saúde
	Gestão do estabelecimento de saúde
	Código do Distrito Administrativo
	Distrito Administrativo
	Subprefeitura
	Supervisão Técnica de Saúde
	Coordenadoria Regional de Saúde
Relação Paciente X Estabelecimento	Caráter da internação
	Competência (AAAAMM)
	Data de emissão (AAAAMMDD)
	Data de internação do paciente (AAAAMMDD)
	Data de saída do paciente (AAAAMMDD)
	Complexidade da internação
	Código da especialidade do leito
	Grupo do procedimento autorizado
	Código do diagnóstico principal
	Código do diagnóstico secundário 1
	Código do diagnóstico secundário 2
	Total geral de diárias
	Diárias de unidade de tratamento intensiva (UTI)
	Diárias de unidade intermediária (UI)
	Dias de permanência
	Tipo de financiamento
	Valor da Parcela